Network Optimization Focus Group (NetOp-FG)

Assessment and Recommendations

September 2011

ATIS is a technical planning and standards development organization that is committed to rapidly developing and promoting technical and operations standards for the communications and related information technologies industry worldwide using a pragmatic, flexible and open approach. More than 350 communications companies are active in ATIS' industry committees, and its Incubator Solutions Program.

< http://www.atis.org/>

The ATIS Network Optimization Focus Group (NetOp-FG), Assessment and Recommendations is an ***ATIS Work Plan*** developed for the **Technical and Operations (TOPS) Council**.

This document is subject to change. This document is a product of the NetOp-FG, and represents the consensus view of the NetOp-FG members. However, nothing contained herein is attributable to any particular member of the NetOp-FG.

Published by
Alliance for Telecommunications Industry Solutions
1200 G Street, NW, Suite 500
Washington, DC 20005

Copyright © 2011 by Alliance for Telecommunications Industry Solutions
All rights reserved.

No part of this publication may be reproduced in any form, in an electronic retrieval system or otherwise, without the prior written permission of the publisher. For information contact ATIS at 202.628.6380. ATIS is online at
< http://www.atis.org >.

Printed in the United States of America.

Table of Contents

EXECUTIVE SUMMARY **5**
1.1 PROBLEM STATEMENT 5
1.2 NETWORK OPTIMIZATION AS A SOLUTION 5
1.3 SCOPE OF EFFORT 5
1.4 ASSESSMENT AND CONCLUSIONS 5
1.5 RECOMMENDATIONS 7
2 INTRODUCTION **7**
3 DEFINITIONS **9**
3.1 SERVICE/NETWORK DOMAINS AND LOGICAL INTERFACES 9
3.1.1 Logical Domains *10*
3.1.2 Interfaces *11*
4 USE CASE OVERVIEW **11**
4.1 CONGESTION-AWARE FAIRNESS 12
4.1.1 Congestion-Aware Fairness based on usage *13*
4.1.2 Congestion-Aware Fairness based on ConEx *14*
4.2 SUBSCRIBER-/APPLICATION-AWARE NETWORK OPTIMIZATION 15
4.3 NETWORK-AWARE SCHEDULING OF CONTENT 16
4.4 USER RATE PLANS 17
4.5 REASONABLE NETWORK PROTECTION & MANAGEMENT 18
4.6 LOAD- AND POLICY-AWARE MULTI-RAN SELECTION 18
4.7 OPTIMIZING USE OF WIRELESS NON-BEARER RESOURCES 19
5 ARCHITECTURAL FRAMEWORK **20**
5.1 MONITOR/PROBE FUNCTION 21
5.2 ANALYTICS FUNCTION 22
5.3 TRAFFIC DETECTION FUNCTION (TDF) AND OTHER INLINE FUNCTIONS 23
5.4 APPLICATION GATEWAY 23
5.5 POLICY DECISION FUNCTION (PDF) 23
6 USE CASE ANALYSIS **24**
6.1 CONGESTION-AWARE FAIRNESS 24
6.1.1 Service/Business Needs *24*
6.1.2 Implementation Options *27*
6.1.3 SDO/Gap Analysis *31*
6.2 SUBSCRIBER-/APPLICATION-AWARE NETWORK OPTIMIZATION 33
6.2.1 Service/Business Needs *33*
6.2.2 Implementation Options *33*
6.2.3 SDO/Gap Analysis *35*
6.3 NETWORK-AWARE SCHEDULING OF CONTENT 36
6.3.1 Service/Business Needs *36*
6.3.2 Implementation Options *36*
6.3.3 SDO/Gap Analysis *39*
6.4 USER RATE PLANS 39
6.4.1 Service/Business Needs *39*
6.4.2 Implementation Options *40*
6.4.3 SDO/Gap Analysis *41*
6.5 REASONABLE NETWORK PROTECTION & MANAGEMENT 41
6.5.1 Service/Business Requirements *41*
6.5.2 Implementation Options *41*
6.5.3 SDO/Gap Analysis *42*
6.6 LOAD- AND POLICY-AWARE MULTI-RAN SELECTION 43
6.6.1 Service/Business Needs *43*
6.6.2 Implementation Options *43*
6.6.3 SDO/Gap Analysis *44*
6.7 OPTIMIZING USE OF WIRELESS NON-BEARER RESOURCES 45
6.7.1 Service/Business Needs *45*
6.7.2 Implementation Options *47*
6.7.3 SDO/Gap Analysis *49*
7 SDO/GAP ANALYSIS SUMMARY **50**

8 LEGAL/REGULATORY SUMMARY 52
8.1 CONGESTION-AWARE FAIRNESS 54
8.2 SUBSCRIBER-/APPLICATION-AWARE NETWORK OPTIMIZATION 54
8.3 NETWORK-AWARE SCHEDULING OF CONTENT 55
8.4 USER RATE PLANS 56
8.5 REASONABLE NETWORK PROTECTION & MANAGEMENT 56
8.6 LOAD- AND POLICY-AWARE MULTI-RAN SELECTION 56
8.7 OPTIMIZING USE OF WIRELESS NON-BEARER RESOURCES 57
9 CONCLUSIONS & RECOMMENDATIONS 57
9.1 SHORTER-TERM STANDARDS NEEDS 57
9.1.1 ATIS Committees/FGs 57
9.1.2 External SDOs 58
9.2 LONGER-TERM STANDARDS NEEDS 58
APPENDIX A: ACRONYMS 59
APPENDIX B: NETWORK OPTIMIZATION FOCUS GROUP MEMBERS 61

EXECUTIVE SUMMARY

1.1 Problem Statement

It is well known that network data usage for both wireless and wireline networks continues to grow dramatically as subscribers' appetite for bandwidth-hungry applications and content increases. For wireless networks, this trend is compounded by strong growth in data-capable smartphones as well as other mobile internet devices. In the future, the expected very large deployments of Machine-to-Machine (M2M) devices will also place significant stress on the existing network.

1.2 Network Optimization as a Solution

To keep up with demand, networks operators can ill-afford to address the issue by simply throwing more bandwidth at the problem. Networks must also become more efficient and must be optimized to better handle traffic growth while maintaining a high Quality of Experience (QoE) for end-users.

In the context of this report, network optimization refers to dynamic and automated changes in network behavior triggered by network resource utilization events that enable the network to adapt to traffic patterns to optimize network, application or subscriber performance as appropriate. Specific network optimization actions may impact how and when traffic is routed, policed, shaped, prioritized, and may include employing optimization functions such as caching, Transmission Control Protocol (TCP) optimization, signaling proxies, video rate adaptation/compression, or buffer management techniques. Network resource utilization triggers can include the indication of congestion or impending congestion in any part of the network. The network can leverage subscriber awareness in its optimization decision in order to ensure fairness and consistency with subscription parameters. Some optimizations target a broad application class such as video content, or Machine Type Communications (MTC).

1.3 Scope of Effort

For the purposes of this report, the issue of network optimization was examined primarily from a technical perspective. While there are significant legal and regulatory issues pertaining to the issue of network management and optimization, some of which are discussed in Section 8 of this report, the report does not provide legal or regulatory guidance regarding particular implementations of the technical issues discussed.

The scope of this document includes both internet access and managed/special services.

1.4 Assessment and Conclusions

In order to look more closely at specific optimization techniques, seven different use cases have been identified for analysis:

1. Congestion-Aware Fairness
2. Subscriber-/Application-Aware Network Optimization
3. Network-Aware Scheduling of Content
4. User Rate Plans
5. Reasonable Network Protection and Management

6. Load- and Policy-Aware Multi-RAN Selection
7. Optimizing Use of Wireless Non-Bearer Resources

In analyzing the above cases, the Focus Group found a number of areas for further standards development. Specifically:

Congestion Awareness - Providing accurate and timely congestion indications from key congestion points in the network (i.e., radio access network (RAN), shared wireline access, high-concentration aggregation points) as inputs and triggers for dynamic network optimization actions is a common need across many of the use cases analyzed in this study. Additionally, the Internet Engineering Task Force (IETF) Congestion Exposure (ConEx) initiative was found to hold some promise of a more end-to-end, holistic approach to managing congestion.

Policy - In order to make policy decisions that enable a wide range of optimization use cases, the Policy Decision Function (PDF) needs standard ways of obtaining inputs regarding both the state of the network and the state of subscribers on the network. This requires a standard interface to the PDF (Policy and Charging Rules Function, or PCRF), providing this information from other systems such as analytics and/or Operations, Administration, and Maintenance (OA&M) systems.

Traffic Detection - Traffic Detection Functions (TDFs) classify traffic in-line and can also apply optimization enforcement actions selectively to the various classified traffic flows. These enforcement actions can take many forms, and it may be impractical to standardize every specific enforcement action and the explicit control of those functions over the Gx interface from the PDF to the TDF. A practical approach would be to provide an application-specific user priority (or optimization index) over the Gx/Sd interface, which would allow operator-defined optimization enforcement actions (coupled to corresponding operator-defined policy rules) to be signaled between these two entities.

Application/Network Interfaces - Several optimization use case implementations use network Application Programming Interfaces (APIs) to interact between network control and applications and devices. One of the most important examples is the use of network APIs for minimizing the signaling impact of massive scale M2M applications. Standard APIs need to be developed that will either replace or synchronize short duration but frequent signaling activity, such as heartbeat methods of determining device connection and health. Another specific example is the need for extending interface definitions on the Rx interface between PDF and Application Gateway to provide a minimum re-try interval, as a way of deferring requested connection services, and avoiding APIs being overloaded by rapid re-requests when the network is too congested to satisfy the request.

Access Selection – In order to support better coordination with policy and charging control subscriber policies, we recommend the use of user data convergence (UDC) for Access Network Discovery and Selection Function (ANDSF) policies. Additionally, ANDSF policies should allow the user equipment (UE) to utilize access network availability information exposed via congestion awareness techniques. Specifically, there is an identified need for a mechanism to inform UE of coarse-grained access network availability at the time of attachment.

In the United States, there are important evolving regulatory requirements in the area of net neutrality to preserve Internet freedom and openness. This report addresses how the defined network optimization framework relates to such regulations in the U.S. After careful consideration, we believe that the methods addressed here can be made consistent with net

neutrality rules. Furthermore, we believe the underlying structure of these optimization use cases, which are built upon the 3rd Generation Partnership Project (3GPP) policy and charging control (PCC) architecture, will also provide great utility for users and network service providers globally. It is very important that companies looking to implement network optimization methods seek guidance from their own regulatory counsels before implementing any mechanisms.

1.5 Recommendations

It is recommended that member companies take leadership roles in drafting contributions and rallying support behind proposals to address the identified standards development needs and the other specific items outlined above and in Section 7 of this document.

Specific actions with the applicable Standards Development Organizations (SDOs) include:

- ATIS - coordinate analysis and align recommendations on points of intersection between the NetOp-FG and other ATIS committees and focus groups - namely, the Packet Technologies and Systems Committee (PTSC), Cloud Services Forum (CSF) and M2M Focus Group.
- 3GPP – in addition to the standards work noted in Section 7, specific focus will be needed to progress the concept of exposing congestion information in the network and use of network congestion information in policy decisions (e.g., PCC and/or ANDSF) for optimization purposes.
- Open Mobile Alliance (OMA) - further work in applicable SDOs such as the OMA to detail the API functional needs of important network optimization use cases.
- IETF - In the longer term, holistic congestion management approaches such as the IETF ConEx approach hold great promise for aligning application needs, user priorities, and network charging, policing, and overall congestion management in a comprehensive way. The NetOp-FG recommends that ATIS member companies actively encourage and advance the ConEx concept, especially in the areas specific to RANs, as well as consistent and practical end-to-end congestion volume-based incentive/enforcement regimes.

2 INTRODUCTION

Network optimization, as used in this report, is defined as any mechanism used to modify network behavior to better meet customer/end-user service quality needs given the available network capacity. The mechanisms identified in this report for network optimization can be applied to fixed broadband, mobile broadband, or managed networks and services. Optimization methods such as the network optimization techniques described in this report are common to all complex systems where the balance of customer service and capacity availability must be considered (such as electrical distribution/transmission systems, transportation systems, and shipping/delivery systems). The telecommunications industry in particular has a long history of providing optimization techniques in a balanced way, such as circuit switched traffic engineering. While network optimization is not new, finding new ways to optimize the network is even more critical in today's telecommunications environment with the advent of smartphones, greater network integration, and dramatically increased broadband data volumes.

Network optimization affects network behavior by changing how traffic is routed, policed, shaped, prioritized, or otherwise optimized. Optimization can be triggered by network events such as congestion or anticipation of congestion by the subscriber/application interactions. Network optimization can also provide the reasonable network management necessary to support a more open and fair broadband internet, and thus a healthier and growing ecosystem.

The high-level benefits of network optimization are:

- Increased user QoE,
- Improved network efficiency,
- Greater flexibility for simultaneously balancing user QoE and network efficiency, and
- Optimal and sustainable investment in broadband services.

The purpose of this report is to describe network optimization in terms of an architectural framework, definitions, use cases and scenarios, requirements, and analysis, and to provide recommendations based on this (e.g., further standards work).

QoE enhancements could be on a per-user or per-application basis, or in aggregate (i.e., the aggregate QoE is improved even though the QoE for a specific subscriber may be adversely impacted). QoE improvements can be measured through customer satisfaction surveys or churn rate metrics. Service providers, network providers, and users all benefit from optimizations that improve application performance, as well as from the resulting business opportunities. Decreased cost can be achieved through better network asset utilization (as a result of network optimization or lower churn rates) as well as reduced operational expense (OPEX).

We believe that network optimization opportunities exist for both wireline and wireless networks. It is worth noting that while these two network types share some aspects in common (e.g., they both involve substantial wireline assets), they also have important differences. For example, wireless bandwidth typically has a higher cost of deployment than wireline, thus making the motivation for and financial justification of optimization much stronger. Wireless also has two fundamentally different characteristics arising from (1) high user mobility and (2) air interface resources that must be managed (and optimized), which both impact optimization greatly. User mobility implies that the user's network access attachment point is frequently changing, and the air interface resources utilized by a given user flow need not, and frequently will not, correlate well with the corresponding resources on the network side. This difference originates from the unique characteristics of the air interface, such as user proximity to the basestation (near vs. far), user mobility and propagation characteristics, and the resulting modulation and coding scheme. The air interface is the source of most user-plane congestion, and its share is rising.

By its very nature, network optimization requires interaction between various network elements and functions to implement any specific optimization. Of course, many optimization algorithms already exist within individual network elements or layers (e.g., air interface scheduling or TCP). Hence interactions between independent optimization algorithms can occur (e.g., from different optimization use cases, or coexistence with already existing algorithms). To resolve these conflicts, it is useful to have a common standardized framework for optimization to insure the stability of the network and the user experience.

Many SDOs have considered the need for a PCC framework. For example, 3GPP has developed a standardized PCC architecture which can be effectively leveraged for end-to-end network optimization. We find this PCC architecture to have considerable merits and is well suited for the use cases presented in this document. Therefore, we believe that evolving network optimization scenarios will require the network to support policy decision and enforcement capabilities. We also believe that certain extensions of network functionality will enable PCC-based optimizations to provide even greater benefits.

Many optimization scenarios require some level of interworking or cooperation between the network service provider and one or more application providers. In some cases, network brokering or interconnect services may be useful to ensure that an application provider can work seamlessly with multiple network providers.

The organization of this document is as follows. Since we discuss network optimization for both wireline and wireless networks, (which often use different terminology,) Section 3 introduces the terminology we shall use for service and network domains and logical interfaces. Section 4 then presents high-level overviews of seven network optimization use cases that were judged most important. Section 5 describes the common architectural framework we use to address the various optimization scenarios, including some new optimization-related elements. Section 6 presents a more detailed user case analysis, noting key service capabilities and implementation options, as well as providing SDO and gap analyses. Section 7 provides an SDO and gap analysis summary. Section 8 provides a legal and regulatory summary at the time this document was written, since network optimization (and the seven use cases) must be compatible with the FCC's "open access" and "reasonable network management" requirements. Finally, Section 9 contains our conclusions and recommendations.

3 DEFINITIONS

Defining and implementing the network optimization use cases outlined in this document will require the cooperation of many different systems, some spanning multiple networks. This section establishes a set of common definitions for the provider and consumer domains involved in the services and networks to be optimized, to be used in the use case analyses.

3.1 Service/Network Domains and Logical Interfaces

This document uses terminology for the logical domains of *Content Provider*, *Service Provider*, *Network Provider*, and *Consumer*[1]. These terms identify logical domains as regions of administrative control or ownership. The architecture has to support the existence of multiple entities with different regions of administrative control and ownership interests. Ownership and administrative control are impacted by a variety of factors including prevailing regulatory regimes, competitive commercial environments, and the commercial strategies of the entities involved. As such, ownership or administrative control may be considered arbitrary boundaries within the functional decomposition of any service architecture, not the basis for that decomposition.

[1] Adapted (simplified form) from terminology defined by the ATIS IPTV Interoperability Forum (IIF), in ATIS-0800007, *IPTV High Level Architecture.*

This decomposition, although typical, does not preclude a single administrative entity from implementing more than one logical domain. In addition, the logical interfaces (inter-domain relationships which may necessitate one or more functional interfaces) between the domains are not always linear, as shown in Figure 1 and described below.

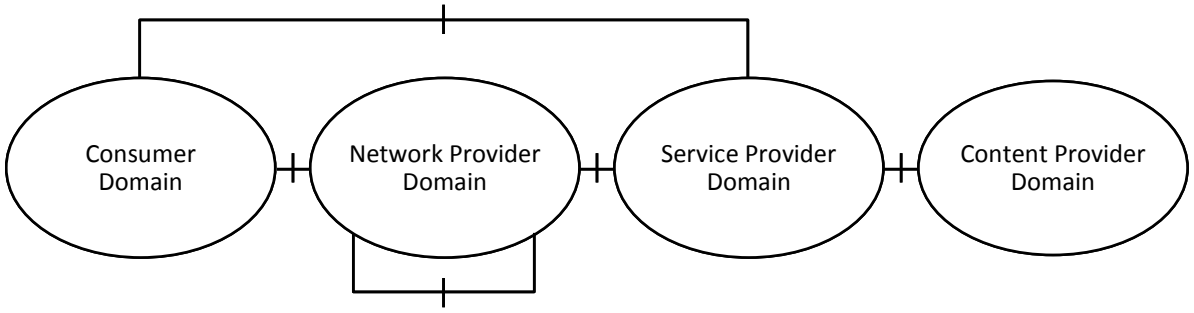

Figure 1: Logical Service/Network Domains and Interfaces

3.1.1 Logical Domains

1. *Consumer*: The domain where the services are consumed. In the consumer domain, a single terminal may be used for service consumption, but a network of terminals and related devices may be used instead. The domain may also be a mobile end device; in this case, the delivery system of a network provider is a wireless Wide Area Network (WAN). This domain is in scope for this analysis.
2. *Network Provider*: The domain connecting customers and service providers. The delivery system is usually composed of access networks and core or backbone networks and uses a variety of network technologies. Traffic traversing this domain may flow through multiple independent operators (e.g., roaming networks or third-party backhaul providers). This domain is in scope for this analysis.
3. *Service Provider:* The domain providing a service to the subscriber. In addition to traditional telco service providers, this domain may include internet application providers, content and application aggregators, or any provider offering services over the network to consumers. This domain is in scope for this analysis.
4. *Content Provider:* The domain that owns or is licensed to sell content or content assets, which may include applications. Although the service provider is the primary source for the customer at home, a direct logical information flow may be set up between content provider and home customer, such as for rights management or protection. This domain, especially acquisition of content by the service provider, is in scope for this analysis. Specifications related to the content provider's content development processes are *not* considered to be in scope at this time.

3.1.2 Interfaces

The following interfaces between domains are of interest to network optimization use cases, and thus are in scope for the analysis of network optimization use cases:

1. *Consumer/Network Provider*: The Consumer/Network Provider logical interface may include data plane traffic delivered to or from the consumer on behalf of the Service Provider, as well as control (e.g., signaling or API interactions) indications between Consumer Domain devices and the Network Provider, e.g., to modify if, when, and how connectivity services are delivered.
2. *Network Provider/Service Provider*: The Network Provider/Service Provider logical interface may include both data plane traffic (e.g., content or service traffic that is processed and/or stored in the network), as well as control (e.g., signaling or API interactions) indications between the Service Provider and Network Provider domains, (e.g., to control if, when, and how connectivity services are delivered over the network.
3. *Service Provider/Content Provider:* The Service Provider/Content Provider logical interface may include content transferred between Content Provider to Service Provider, as well as control and management indications between Content Provider and Service Provider domains (e.g., to coordinate content management or control Quality of Service (QoS) for real-time content).
4. *Consumer/Service Provider*: The Consumer/Service Provider logical interface may include data plane traffic like content or interactive service traffic, which transits the Network Provider domain in between the Service Provider and Consumer domains, and may also include control (e.g., signaling or API interactions) indications between Service Provider and Consumer domains (e.g., to request or modify services and to determine if, when, and how they are delivered).
5. *Network/Network Provider*: When multiple Network Providers are involved in the delivery of a service, the Network/Network Provider logical interface may include data plane traffic (e.g., content or service traffic processed and/or stored in the two networks), as well as control (e.g., signaling or API interactions) indications between two networks, e.g., to control QoS, multicast, or other aspects of if, when, and how connectivity services are delivered across the two networks.

While other inter-domain interfaces may be possible, they are not deemed to be of primary concern to the types of network optimization use cases being studied, and are considered out of scope for this analysis.

4 USE CASE OVERVIEW

This section provides a description of each of the network optimization use cases analyzed in this study. Note that the use cases outlined here are not mutually exclusive; many of them can be active in a network concurrently, and aspects of them can be combined to form more complex use cases. While most of the use cases in this document deal with optimization of user data plane traffic, two areas—Section 4.7 (Optimizing Use of Wireless Non-Bearer Resources and aspects Section 4.5 (Reasonable Network Protection & Management)—concern optimization of the signaling plane for wireless networks. Some of the optimization techniques identified in the following use cases may not be needed for some networks with very high bandwidth, such as fiber access.

4.1 Congestion-Aware Fairness

Congestion-Aware Fairness involves optimization of network capacity usage to mitigate the impact of congestion on users during peak periods. The use case breaks down into two scenarios. The first scenario, which is implementable in the near term, uses existing mechanisms to identify heavy users and to enforce policy during peak periods of congestion. The second scenario assumes the adoption of a ConEx mechanism, which provides inline, real-time indications of which flows and users are contributing to congestion. The two scenarios are described in their respective sections below.

The scenarios focus upon the case where user-plane traffic congestion[2] occurs within the same QoS/QoS Class Identifier (QCI) class. Handling congestion across multiple QoS levels has already been addressed by 3GPP and other standards bodies.

We begin with the observation that, depending on type, networks presently have one or more specific gaps to address in congestion awareness.

The first gap, which applies to all types of networks, is that most existing network fairness mechanisms (for example, weighted scheduling or proportional fairness algorithms) do not provide feedback between a user's contribution to network congestion and the treatment applied to that user's traffic at times of network congestion. As a result, there is no incentive for users to minimize or avoid congestion. Within a given QoS (or QCI) class, users who generate large volumes of data during times of congestion receive the same relative bandwidth fraction and priority as users who rarely contribute to congestion.

By using metrics that are correlated with network congestion and applying enforcement or incentive actions consistent with those metrics, congestion-aware fairness allows higher levels of average network utilization while preserving users' QoE. As examples that could be used singly or in combination, heavy users may be encouraged to avoid congestion or may face de-prioritization of their traffic during times of congestion.

The second gap, which is specific to mobile networks, is retaining user awareness on intermediate time scales of minutes or hours, so that one can more fairly address resource allocation between users with very different data rates and user session activity. Ordinarily, one anticipates that a classical fairness mechanism working at short times will also address fairness on longer time scales. However, this is not always true for mobility cases with at least one of the following characteristics:

- There are a mix of users with widely varying session activity behavior, such as users with long active sessions (e.g., file transfer protocol (FTP) or streaming) together with "always on" users with frequent but typically brief session activity.
- The short-time fairness mechanism discards the user state when they transition to the idle state.

Retaining user awareness on intermediate times poses special challenges for mobile networks, as one must retain the user's context as they move through the network as well as over periods

[2] In addition, a network optimization use case is possible for addressing signaling-plane congestion arising from always on or "chatty" data applications.

of time when they transition between active and idle states. Today, key "classic" fairness mechanisms within mobility standards are non-stateful and act only on short times.

A third gap, again specific to wireless networks, concerns the need to expose additional information to improve the effectiveness of congestion-aware mechanisms. Because of the broad range of wireless link quality that exists for different users, air interface utilization need not correlate well with throughput as measured within the transport or backhaul network. For example, within Long Term Evolution (LTE) two users in the same cell could have ≥25-fold differing bitrates[3] while using the same number of air interface resource blocks. Hence another goal for wireless networks would be for the RAN to make available[4] some measure of its users' air interface resource utilization (e.g., transmit power, Walsh codes, or resource block utilization) averaged over some coarse time scale (e.g., 10 seconds – 1 minute). This information would improve the input data quality of resource utilization when the air interface is the relevant congestion point, providing more informed policy-based decisions and enforcement. However, enforcement or other policy actions based on this information should be tempered by the fact that radio resource utilization is not subject to user visibility or control, and that users with poor wireless link quality may already be penalized by proportional scheduling algorithms.

The congestion-aware fairness methods described below represent an important new category of reasonable network management because they benefit both users and operators. For even the heaviest users, these methods help make unrestricted access to bandwidth under normal network conditions more practical and affordable, while ensuring that the majority of users are not heavily impacted by the consumption habits of the consistently heaviest users. Service providers need these types of network optimization tools to ensure that access to network resources - especially during congestion peaks - is managed fairly for all users.

4.1.1 Congestion-Aware Fairness based on usage

This scenario can be viewed as a variant of the Comcast "fair share" algorithm[5] for managing "extended high consumption users". Hence, there are three general functions associated with the scenario:

1. Measurement of congestion levels in the network;
2. Measurement of subscriber usage; and
3. Policy-based decision and enforcement during peak periods of congestion based upon data gathered via the first two functions.

The scenario considers network optimization using two levels of complexity. In the first level, subscriber's longer term network usage is measured independent of congestion levels in the network. In the second level, network congestion levels are applied to the measurement of per-subscriber usage to correlate that usage with the frequency and/or severity of network

[3] This comes from comparing the maximum and minimum number of bits possible per resource block, as defined by the 3GPP LTE modulation and coding schemes. Even greater differences are possible if one includes MIMO.

[4] To other network elements and/or devices.

[5] See IETF RFC 6057 Comcast's Protocol-Agnostic Congestion Management System.

congestion. In both levels, the potential approaches include triggering enforcement or incentive actions to mitigate such congestion.

Via the methods described in Section 6, the network can accurately monitor and flexibly threshold congestion levels at relevant points in the network. Similar methods can be used to monitor usage per subscriber. The per-subscriber monitoring might take a number of forms, including but not limited to:

- Average or total usage during a sliding window, with a measurement interval in the range of minutes or hours (although it could be defined over a longer period). This represents Level 1 monitoring per the above description.
- Average or total usage only during periods of congestion. This represents Level 2 monitoring.
- Average or total usage which is weighted/valued differently during periods of congestion than during non-congested periods. This represents Level 2 monitoring.

The network- and user-monitoring data can be used by policy controls to trigger enforcement or incentive actions which limit or de-prioritize the sustained heavy users' consumption of network capacity during times when the network is congested. This contributes to greater long-time "fairness" measured across users of widely varying bandwidth and session activity, moderating the adverse QoE impacts that consistently heavy users have upon the QoE of other users (e.g., browsing responsiveness, email/page/file download times, and video performance).

Policy controls allow for enforcement and/or incentive actions to factor in the subscriber's service tier, roaming status, usage vs. monthly cap, location (geographic, cell type, or specific cell), and possibly even historical (much longer term) usage patterns. This provides operators significantly improved flexibility to manage their networks in order to better tailor their range of service offerings to meet or even improve the QoE of their users.

4.1.2 Congestion-Aware Fairness based on ConEx

A long term goal may be to adopt ConEx (described in Section 6) as an enabling mechanism for determination of congestion, as well as a component of the mechanism for responding to such congestion.

ConEx enables the same charging and other policy mechanisms described in the above section, but with more precision, since they can be based directly on the congestion to which a flow (or the subscriber associated with a set of flows) contributes, rather than on an indirect parameter such as usage. ConEx also enables additional policy mechanisms. For example, congestion can be measured against a monthly allowance of congestion credits, in which case traffic that avoids congestion (by virtue of being sent at non-peak periods or by using a "scavenger" protocol such as Low Extra Delay Background Transport (LEDBAT)) does not count against the bank of credits. In another example, a policing function could work as a "congestion policer," policing only that traffic which contributes to congestion.[6]

With regard to the hosts generating such flows, direct exposure of congestion motivates the use of scavenger protocols for non-time-sensitive traffic, since traffic that actively avoids congestion

[6] See "Congestion Exposure (ConEx) Concepts and Abstract Mechanism," draft-ietf-conex-abstract-mech-01, p. 13.

will now have a lower cost associated with it. Currently there is a disincentive to use scavenger protocols because they yield bandwidth disproportionately to other flows without realizing any corresponding benefit to the user.

Given the early state of ConEx development in the IETF, it is unclear when or even if this scenario will be feasible. However, the potential benefits from using a ConEx mechanism are such that it is advisable to track the progress of the IETF work and possibly influence its direction via liaisons or direct participation.

4.2 Subscriber-/Application-Aware Network Optimization

This scenario pertains to the set of network optimization uses cases that give the subscriber or the application provider the option of easily selecting how the network should treat different applications or application classes for a specific subscriber. Subject to net neutrality considerations, the network may also use application awareness to optimize reasonable network management.

For this scenario, the network detects that a flow or traffic pattern exists for a specific subscriber and/or application and optimizes that flow depending on network, subscriber, and application policies. The network may become aware of a specific flow to be optimized through a variety of means. For example:

- The network may detect a specific flow via a network based packet flow detection function.
- A specific flow descriptor may be signaled to the network from an application server or from the application software running on the subscriber device.

The specific optimization applied to the flow may include well known QoS policy treatments (e.g., using the IETF Differentiated Services (DiffServ) model) or may affect the flow transport/application protocol and/or packet contents. For example, it is common in wireless networks to implement traffic optimization operations that do the following:

- Improve TCP performance with lossy, high delay wireless networks;
- Reduce bandwidth consumption by providing additional compression (either loss-less or lossy depending upon media type);
- Improve application performance through application specific protocol optimizations;
- Improve perceived performance with screen size specific optimizations; and
- Improve bandwidth utilization and/or user experience for video content through rate adaptation, caching and delivery optimizations.

The network optimization could be applied on a per-subscriber basis, a per-application basis, or a combination of subscriber/application. For example:

- The optimization may be applied to all flows (within a specific service classification) of a specific subscriber based on a subscriber profile and policy.
- The optimization could be enabled for all subscribers using a specific type of application, such as IP (Internet Protocol) multimedia subsystem (IMS) services, video optimization, or parental controls.
- The optimization could be enabled for a specific subscriber when running a specific type or category of application.

The optimization event may trigger charging for the subscriber, the application provider, or both.

Subscriber-/Application-aware network optimization use cases allow more effective network management and provide opportunities for differentiation by giving customers more control over their broadband experience. For example, quality streaming of a High Definition (HD) sports channel could be bundled into a premium subscription tier.

For the subscriber, subscriber-/application-aware network optimizations can do the following:

- Improve the QoE through increased application response times (e.g., faster page load or faster downloads) or improvements in perceived quality of media based applications (e.g., video and/or audio streaming); and
- Enable access to new applications that otherwise would not be feasible without network optimization.

For the service provider(s), these optimizations can do the following:

- Lower network costs by better utilizing network assets;
- Increase overall user satisfaction; and
- Increase revenue by enabling new and/or differentiated service capabilities to both the subscriber and the application provider.

4.3 Network-Aware Scheduling of Content

This scenario seeks to provide more uniform network utilization by delaying less time-sensitive traffic to later times when the network load is lower. This scenario can additionally be used in combination with the Load- and Policy-Aware Multi-RAN Selection Scenario to create more complex optimization environments. For example, when multiple access technologies are involved (e.g., Wideband Code Division Multiple Access (W-CDMA), LTE, CDMA, Wi-Fi, femtocell, wireline, etc.), one can first assess whether there is an available access technology that is currently uncongested and use that network. If all available access networks have some level of congestion, time shifting methods can be applied when appropriate to defer the data transfer at to a later time. This cycle can be repeated as necessary. By reducing network congestion during peak periods, this approach can simultaneously provide better QoE to users through reduced congestion and fewer video halts, while improving average network utilization and thus lowering total network investment for the service provider.

The network optimization could also make use of service differentiation that varies according to the subscriber's user class, application, QoS, device, and/or additional factors.

For this scenario, the main enabling factors are:

1. Awareness of the presence or absence of network congestion on the presently serving access network.
2. User willingness to accept time-delayed delivery of content, potentially via incentives.

Network congestion can be inferred by many means. These mechanisms are discussed in more detail in section 6 of this document.

User acceptance of time deferred content likely requires an appropriate business and/or content delivery model that motivates the user to allow and even expect delayed delivery of certain content. Some possible examples might be:

- Rate plans that have a separate treatment for time-shifted content (e.g., with additional data usage limits and/or discounts).

- Purchased content (movies, eBooks, applications, etc.) which specifically call for time-shifted delivery, potentially with discounted delivery and/or a different data usage classification.

On the user expectations front, content time shifting needs to be executed with minimal impact on users but with a realistic view of implied service "guarantees". We note that it is not possible to rigorously guarantee content delivery at a later time, as the user could have no/poor connection capabilities at later times. Hence it appears best to start with non-real time content whose large size makes it very difficult to deliver (or even precludes delivery) under present "loaded" network conditions. The type of content that seems natural for this approach can either be large on an individual user basis (e.g., large files and flows such as movies and large video files, eBooks, large applications, or software/firmware updates) or in an aggregate basis (for instance, when dealing with huge numbers of devices with comparatively small messages, this could also be relevant to machine-to-machine communications).

4.4 User Rate Plans

The use of different tiers of user rate plans is also a tool for optimizing usage of network resources. There are many possible forms of rate plan differentiation, many of which are in broad use by networks today. Generally they seek to restrain excessive usage of the network and to provide better alignment between what users pay and the level of their network usage.

Examples of user rate plan variations include:

- Rate Capping - The user's uplink and downlink traffic is limited to a maximum bit rate, which differs according to the service tier to which the user is subscribed.
- Usage Capping - The user's total usage during a billing period is limited to a maximum volume (measured in MBs or GBs, and depending on the service tier), after which the service is either blocked or rate-limited. The user is often offered the option to upgrade immediately to a higher service tier.
- Overage Pricing - Similar to the usage cap, the user has a given tier-specific usage volume which is bundled into the price of each service tier, and usage above that threshold is charged at an overage rate which is typically higher than the base bundle.
- Time-dependent Pricing - Used in conjunction with usage caps or overage pricing, users can be encouraged to shift their consumption habits toward off-peak periods by discounting or zero-rating their usage during these fixed (a particular time of day) or dynamically determined periods. These time-shifting aspects are covered in the "Network-Aware Scheduling of Content" use case area introduced in Section 4.3.
- Application Zero-rating - Premium subscription tiers can include free access to specific bundled applications or content offered by the operator or a third party, which is part of the value to the user of subscribing to a higher service tier.
- Congestion-sensitive Pricing - Longer term, it may be possible to use congestion exposure information to align what a user pays with the volume of congestion they cause in the network rather than simply their usage volume. In combination with enhanced congestion management protocol capabilities in the end-systems, users

and their applications can be incentivized to reduce the load that lower priority traffic places on the network during periods of congestion.

Implementation of tiered user rate plans often involves policy control to provide flexibility in defining/modifying and executing enforcement of rate plan variations, and charging systems to account for usage.

4.5 Reasonable Network Protection & Management

In this scenario, network optimization mechanisms are used to provide protection for the network and its users from harmful traffic (e.g., denial of service (DoS) attacks, malware, rogue user behavior, etc.) and illegal traffic.

This network optimization provides a variety of benefits including:

- Increased network security and reliability by reducing DoS attacks and the proliferation of malware and infected software.
- Reduced network congestion by removing harmful traffic from the network
- Increased user and business/enterprise security by reducing the proliferation of vulnerabilities.

Harmful traffic may be targeted toward users or network elements such as network servers, Domain Name System (DNS) servers or other network signaling and routing elements.

Typical implementation approaches will use either network or device client-based detection capabilities to invoke network policies to quarantine and/or throttle the impacted flows until the issue has been resolved. Upon detection of harmful traffic, the subscriber associated with the originating harmful traffic should be notified if possible.

4.6 Load- and Policy-Aware Multi-RAN Selection

Today's wireless service provider networks and wireless devices are increasingly multi-technology. Now add to this Wi-Fi access via multiple network providers—at home, work, and in public areas. At any given location and time, service availability and the state of congestion will in general vary for each access network and technology. While such situations in principle present the user with access diversity, (a good thing,) determining and enabling the optimal access technology is often a painful and manual process.

This scenario specifically addresses the use of logic in the device or client, and/or the network(s), such that the device could either determine or be informed as to which RAN (e.g., 3G/4G or Wi-Fi) and associated radio frequency (RF) carrier or band would be most suitable.

Selection criteria could include:

- The present state of the available access networks, potentially including information such as congestion, and/or available throughput-related measures for both the wireless air interface, along with associated wireline backhaul (e.g., Wi-Fi, femto, and/or cellular) for each access technology and carrier/band;
- The specific user application (including QoS and other service requirements);
- Associated service costs;
- User preference(s); and

- Device characteristics (e.g., battery condition).

The resulting benefits from this capability may include:

- More efficient and fairer utilization of air interface and backhaul resources for the various wireless networks (benefitting wireless service providers); and
- Improvements in typical user experience, as well as potential reductions in data surcharges (benefitting users).

In order to affect this optimization (and to avoid the likely negative impacts associated with other schemes involving active network service performance "probes" initiated by the device), one would likely want the networks to monitor and potentially share (or broadcast) system load information on some coarse time scale (seconds or longer).

We also note that network-controlled schemes already exist for certain inter-carrier/inter-technology selection and/or load balancing. However, such schemes are usually not widely deployed for the following situations:

- Across different technology families (e.g., between CDMA and High Speed Packet Access (HSPA), or cellular and Wi-Fi, etc.);
- Between networks built by different infrastructure vendors; and
- Between networks owned by different service providers, enterprises and/or operators.

4.7 Optimizing Use of Wireless Non-Bearer Resources

This scenario applies to congestion not on the user plane but on the signaling plane. This may occur in environments with a large number of chatty smartphone applications or machine-to-machine applications. In these cases, the amount of data transmitted may be small, but the signaling traffic such as connection setup, teardown, or handover can be significant. This type of load occurs predominately in wireless networks. Due to the critical role of signaling traffic, the signaling congestion and overload can cause serious network performance downgrade or even shutdown due to limited network resource in the wireless network. It is therefore important to consider optimization for non-bearer resources.

This differs from many of the previous scenarios in that centralized policing is inadequate. Since the access network is the critical resource in many cases, the accesses to this network must be prohibited or minimized at an early stage while or before the signaling congestion or overload occurs.

Some of the specific types of usage that can drive signaling congestion are:

- Access by low priority machine type devices when the network is overloaded or nearing overload;
- Large numbers of devices acting synchronously and all trying to communicate at the same time;
- Chatty applications that send small amounts of data frequently (status updates or keep-alive messages);
- Devices performing unnecessary signaling (e.g., mobility signaling for devices that don't move or CS signaling for data-only devices);

- Devices moving too aggressively between different access technologies or networks; and
- Temporary network maintenance triggering massive application or device re-registration.

Since these optimizations must occur early and often require preventative measures and indications because of the high possibility of massive chain reactions in the network, these changes are typically required in the signaling flows themselves.

In many cases, the devices and applications' network-unfriendly behaviors are the main cause of network signaling congestion and signaling traffic inefficiency; therefore, device and application optimization also needs to be considered for the E2E signaling optimization solution, for example devices strictly following the back-off time for the retries or devices implementing a dormant optimization to reduce the signaling traffic for frequent active/sleep switching.

5 ARCHITECTURAL FRAMEWORK

The scenarios described above engage many different elements and functions in the network. Optimization typically requires interaction between these network elements and functions. Figure 5.1 below describes the common architectural framework used in this document as different approaches and potential solutions are outlined to address each of the optimization scenarios considered. This generic network framework introduces a number of new optimization-related elements that may not have well-known names or definitions within existing standards, yet may be significant to the SDO and gap analysis. These functions include:

- Monitor/Probe Function
- Analytics Function
- TDF – (newly defined in 3GPP)
- Application Gateway

These functions are described in more detail in the following sections.

In addition, the architectural framework assumes a common, UDC-based environment for user data. The UDC concept (3GPP TS 22.101) supports a layered architecture, separating the data from the application logic so that user data is stored in a logically unique repository allowing access from core and service layer entities called Application Front Ends. User data convergence is an optional concept to ensure data consistency and simplify creation of new services by providing easy access to the user data, as well as to ensure the consistency of storage and data models and to have minimum impact on traffic mechanisms, reference points, and protocols of network elements.

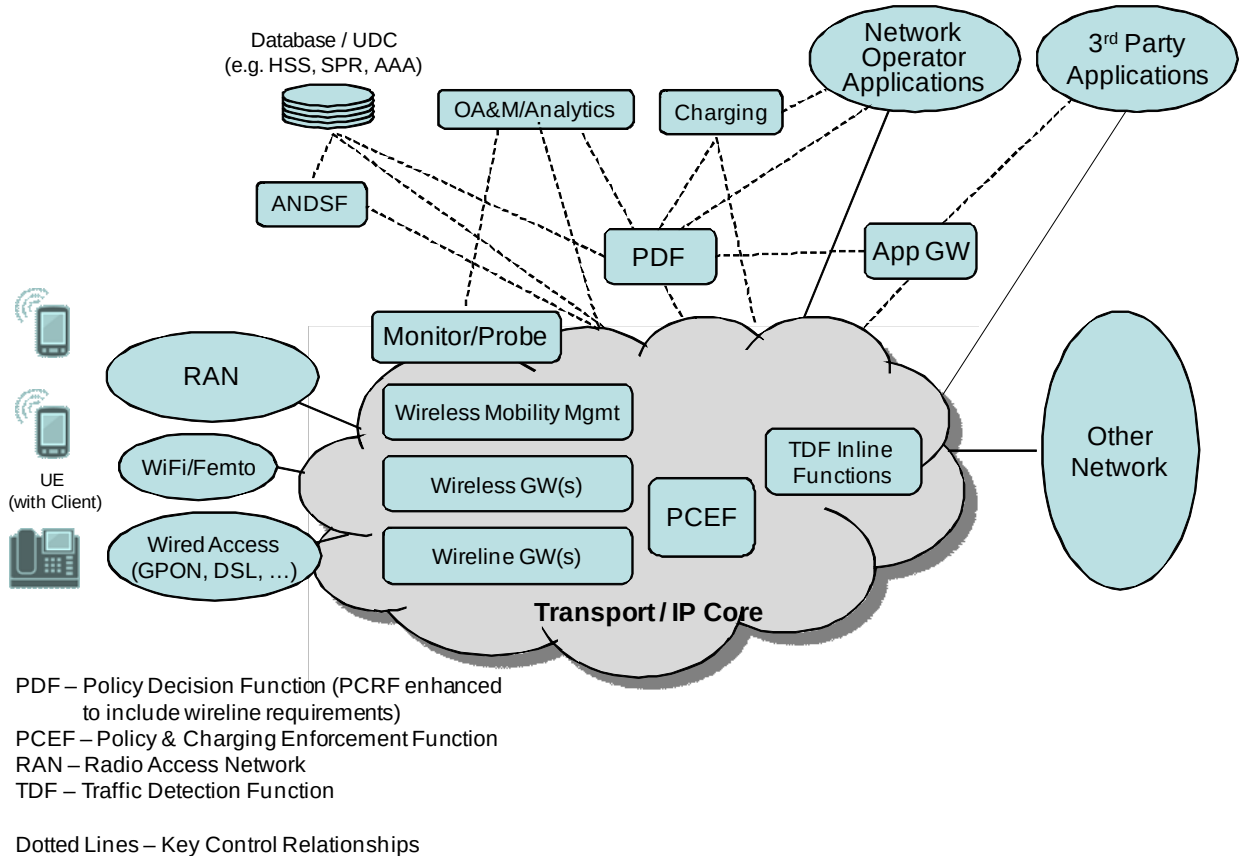

Figure 5.1 – Network Optimization Architectural Framework

5.1 Monitor/Probe Function

Many of the network optimization scenarios assume that some measure of congestion along with other aggregate measures of user traffic behavior are available as input to the optimization decision. A common way to gather this type of information today is to utilize monitoring techniques where user plane traffic along with control/signaling traffic has the following qualities:

- Non-intrusively monitored at various points in the network;
- Summarized; and
- Forwarded to an analytics function to perform an analysis on the data to extract some relevant information or statistics that can be used in the optimization process.

Monitoring can be done through various means leveraging mechanisms such as:

- Router port mirroring;
- Log files or other management related data collection capabilities built into functional elements in the network; and
- Specific probes or devices that may pre-process data to be monitored.

By way of example, the monitor function will often reduce the data set by sampling and/or stripping out the packet payload while preserving the headers. Congestion can be detected by

observing TCP sync events based on TCP/IP header information. The TCP sync analysis could be done by the analytics function. Monitors can also report traditional measures of congestion such as packet/bit flow rates, percent link level utilization, and other measureable traffic attributes.

Monitor functions are commonly differentiated from other traffic detection functions (such as Deep Packet Inspection (DPI)) by the following characteristics:

- Monitoring activities are non-intrusive, with no impact on network capacity and performance;
- Monitoring activities are often applied to both user and control plane traffic to enable control/signaling events and to allow multiple user plane traffic flows to be correlated (via analytics); and
- Monitoring activities are not real-time functions (though they may be nearly real-time). Rather, they provide a summary over time (for example, average congestion or traffic usage over a five-minute interval).

Active probes can also be used to assess network congestion and performance. These probes actively insert small amounts of data into the network and extrapolate network performance based on the treatment of this inserted flow. An example of an active probe is the well known ping capability that allows one to assess the reachability of IP hosts and the latency in the network path between these hosts.

Where possible, the network monitor and probe functions should log congestion events such as:

- Time of event (start time and length);
- Affected part of the network; and
- Subscribers contributing to congestion in the affected part of the network during the event.

5.2 Analytics Function

The analytics function is used with the data extracted by the monitor/probe functions, as well as existing OA&M performance data, to generate traffic information. For example:

- TCP sync behavior can be used to imply congestion for a specific flow.
- Correlation of congestion indications across multiple flows can be used to pinpoint the specific network bottleneck. For example, if the specific cell-sector of a flow is known and if a high percentage of flows within that cell sector are in congestion (based on TCP sync indications), one can infer that the cell sector is in congestion.
- Correlation of control/signaling information with traffic flow information can be done to create a network traffic view that is not otherwise possible.

OA&M performance, network utilization, and log data can also be used to pinpoint areas of congestion and provide other specific information that may be helpful in determining appropriate network optimization recommendations.

Once an optimization recommendation is determined, this information can be forwarded to the PDF for a policy decision, if needed. The interface for this exchange is currently not defined by standards.

5.3 Traffic Detection Function (TDF) and Other Inline Functions

The TDF performs application detection and reporting functions on data flows in the network. In addition, the TDF may enforce gating, redirection (traffic steering), and bandwidth policies on flows. In typical operational scenarios, the TDF may be integrated with the policy charging enforcement point (PCEF) and/or may be integrated in other inline flow processing elements. A key set of inline functions applicable to network optimization provides inline flow optimization functions.

In this report, flow optimization refers to a class of optimizations that result in a reduction of bandwidth used by a flow without adversely affecting the user experience. Flow optimization in the industry may also be referred to as "media optimization" or "content optimization".

The Flow Optimization functions may perform one or more flow-specific functions, including:

- TCP optimization (often used in wireless networks to accommodate for high latency, lossy wireless conditions);
- Content compression (lossy → applied to media or lossless → applied to data content);
- Application-specific protocol optimizations (object pre-fetch, etc.); and
- Video optimizations including caching, transrating, transcoding, just-in-time delivery.

The TDF is further described in the ATIS Policy Management Focus Group (PM-FG) Final Report and further standardized in 3GPP TS 23.203.

5.4 Application Gateway

Some of the network optimization scenarios suggest an interface between the network and application providers to exchange information. For example, for Subscriber-/Application-Aware Network Optimization, this interface can be used to allow an application provider to signal a Network Provider, indicating that a specific flow for a specific subscriber should be given differentiated treatment. Alternatively, the interface could be used to enable the Network Provider to expose specific network information (e.g., load/congestion information) to an application provider to enable optimization. The Application Gateway (AppGW) facilitates the application interfaces between service providers and application providers by providing the necessary security functions required to protect the network for these interfaces. The AppGW connects to the PDF using standard PDF application interfaces (such as the Rx interface defined in 3GPP).

5.5 Policy Decision Function (PDF)

The PDF is responsible for evaluating policy rules when interacting with the application and enforcement points to determine converged network resource requirements. The PDF is discussed in some detail in the ATIS PM-FG final report. Different SDOs have standardized policy decision and enforcement functionality using a variety of frameworks. The PM-FG concluded that there is a business driver for developing a converged policy architecture and recommended that policy convergence should take 3GPP PCC as the starting point for convergence.

Since many of the use cases addressed in this report can be applied to both wireless and wireline environments, the term PDF is used to represent a generic policy management

environment. However, in the spirit of convergence, the SDO analysis will focus more specifically on the 3GPP PCC framework where the PDF is embodied within the 3GPP PCRF.

6 USE CASE ANALYSIS

This section provides a more detailed analysis of needed service capabilities, implementation options, and areas for standards development for each of the seven use cases. Up front, a few points common to multiple use cases have been highlighted.

Several of the use cases described in this section require an assessment of network congestion. Network congestion can be inferred in several ways. Simple schemes such as those based upon monitoring TCP performance and retransmissions can be deployed with little or no modification of existing network elements.

More accurate and/or localized congestion monitoring can be supported via a combination of active or passive network monitoring tools, in-line traffic detection functions, charging systems, and devices.

Explicit Congestion Notification (ECN)[7] provides another indication of congestion where that mechanism is supported. ECN provides the benefits of inline, real time congestion information, and of being visible both to the endpoints associated with the congested flow and to network nodes at or downstream of the point of congestion.

A long term goal may be to adopt ConEx as an enabling mechanism for determining congestion. In this future state, a ConEx indicator provides exposure (at the IP layer) of congestion experienced by individual flows along the end-to-end network path. This indicator, which is set by the sending host, provides exposure of congestion from end to end in the transmission path, including to nodes encountered before the point where the congestion occurred. This provides the entire network path with per-flow, real time congestion information which it may use to determine the priority or other actions associated with the flow.

Most of the call flows in this section illustrate wireless networks, but many of them would equally apply to wireline networks.

6.1 Congestion-Aware Fairness

6.1.1 Service/Business Needs

In the analysis of the Congestion-Aware Fairness network optimization scenario, the following service needs and capabilities were considered.

6.1.1.1 Service/Business Needs (non-ConEx)

The following service capabilities apply to Congestion-Aware Fairness without the use of a ConEx mechanism.

SERVICE CAPABILITY: The Network shall be able to detect and quantify user plane congestion events, including determining:

[7] See IETF RFC3168, "The Addition of Explicit Congestion Notification (ECN) to IP."

- Congestion level (binary or multi-level) of the wireless RAN (individual cell-sector), or wireline access network;
- Other potential network bottlenecks such as backhaul, aggregation, or core network congestion;
- Start and end times—transitions above and below congestion threshold(s)—of the congestion level(s); and
- Identification of subscriber flows contributing to the congestion event, specifically the user flows on the congested network segment. Of special importance are the users that have generated a disproportionate share of the congestion.

SERVICE CAPABILITY: The Network shall be able to identify heavy users as defined by:

- The fractional utilization by the identified user as compared to a subscription parameter (e.g., peak downstream bit rate or volume over specified interval); and
- Use over specific duration of time (e.g., over a 15 minute interval).

Both the time period and fractional utilization thresholds shall be settable by the operator based on operator policy.

SERVICE CAPABILITY: The Network should be able to identify heavy users whose usage has contributed to network congestion as defined by:

- The fractional utilization by the identified user as compared to a subscription parameter, correlated with user plane congestion events in the part of the network (cell and/or backhaul, wireline access, or core) impacted by that user;
- Using specified weights such that usage that has contributed to network congestion can be weighted differently from usage that has not contributed to network congestion; and
- Use over specific duration of time (e.g., over a fifteen-minute interval).

The time period, fractional utilization thresholds and weights should be settable by the operator based on operator policy.

SERVICE CAPABILITY: The Network shall be able to correlate information regarding subscribers' thresholded usage level with network congestion in the part of the network (cell and/or backhaul, wireline access, or core) impacted by that subscriber, so that enforcement actions on heavy users can be selectively applied only when that subscriber is using a part of the network in congestion.

SERVICE CAPABILITY: The network should be able to factor in subscriber attributes in determining the desired action, including:

- Subscriber usage level (over a specified period or sliding window);
- Subscriber total usage volume over billing period;
- Subscription tier;
- Subscriber location (geographic, cell);
- Subscriber roaming indication; and
- Subscriber priority or emergency status.

SERVICE CAPABILITY: The network shall be able to apply policies to identified subscriber flows, which may include:

- Policing and/or shaping traffic to a specific maximum data rate;
- Changing QoS parameters and flow priority (e.g., wireless bearer QCI, wireline layer 3/2 priority bits);
- Change charging identifier during the enforcement period; and
- Other incentive actions such as discounted de-prioritization.

6.1.1.2 Service/Business Need (ConEx)

If congestion-based fairness is based on a congestion exposure mechanism, some of the requirements in the previous section may not be necessary. For instance, identification of specific enforcement periods and events may be supplanted by real time, per-flow monitoring of congestion indicators and localized application of policy based on congestion credits vs. debits.

The following service capabilities apply to Congestion-Aware Fairness based on a ConEx mechanism.

SERVICE CAPABILITY: The network and user endpoints shall be able to support a to-be-defined ConEx mechanism, including:

- Indication of whether or not a given flow supports the ConEx mechanism;
- Indication of whether a given flow has experienced congestion; and
- Indication of whether a given flow is building up credit for use against future congestion.

SERVICE CAPABILITY: The network shall be able to mark packets as experiencing congestion consistent with the ConEx mechanism, including:

- ConEx marking by wireline-based nodes (including nodes within the packet core for mobile networks) consistent with the ConEx mechanism; and
- ConEx marking by RAN nodes to reflect radio resource congestion consistent with the ConEx mechanism.

SERVICE CAPABILITY: The Network shall be able to identify users who contribute to congestion as defined by:

- The aggregated congestion volume of the flows associated with that user;
- Use over specific duration of time (up to a billing interval); and
- For mobile networks, congestion volume may also need to be accounted for on a per-Access Point Name (APN) basis.

Both the time period and aggregated congestion thresholds shall be settable by the operator based on operator policy.

SERVICE CAPABILITY: The network shall be able to factor in subscriber attributes in determining the desired action, including:

- Subscriber congestion volume (over a specified period or sliding window);
- Subscriber total usage or congestion volume over billing period;
- Subscription tier;
- Subscriber location (geographic, cell);
- Subscriber roaming indication; and
- Subscriber priority or emergency status.

SERVICE CAPABILITY: The network shall be able to apply policies to identified subscriber flows, which may include:

- Policing and/or shaping traffic to a specific maximum congestion rate;
- Changing QoS parameters and flow priority (e.g., wireless bearer QCI, wireline layer 3/2 priority bits);
- Changing charging identifier during the enforcement period; and
- Other incentive actions such as discounted de-prioritization.

6.1.2 Implementation Options

This use case can be broken into three main actions:

- Detection/thresholding of heavy users;
- Detection/thresholding of network congestion; and
- Enforcement action.

For each of these actions, there are multiple implementation approaches. Figure 6.1.2.1 illustrates the flow for a wireless-network-based Congestion-Aware Fairness use case, showing the use of network monitor and analytics functions or charging systems for the first action, and monitor/analytics for the second action.

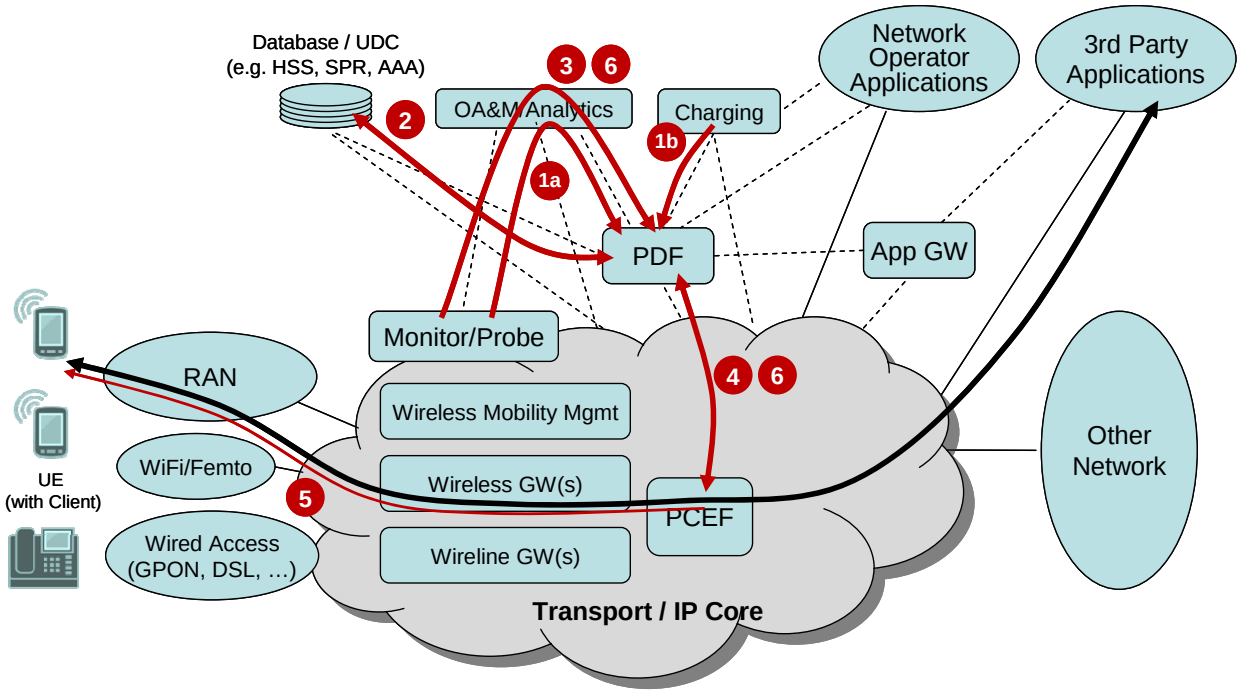

Figure 6.1.2.1-1 Congestion-Aware Fairness

1. Determining which users are considered heavy users at any point in time is based in large part on usage information, which can be obtained via consolidation and thresholding of network monitoring (shown as 1a), or from extraction of information from the Online Charging System (OCS) (shown as 1b), using the newly defined 3GPP Sy interface between the OCS and the PCRF.[8]

2. The PDF uses these usage threshold crossing indications, combined with subscriber profile and other information (service tier, location, roaming status, monthly volume cap status, current services, etc.), to determine if a subscriber is flagged as a heavy user. This state could be retained with the PDF, or set in the Subscriber Profile Repository (SPR) as shown in the diagram.

3. The identification of congestion threshold crossings can be implemented by monitor and analytics functions (as shown), and/or by load indications sourced from Network Elements (NEs) and consolidated by OA&M systems.

[8] Longer-time stateful awareness of users can be implemented at multiple mobile network locations. The basestation scheduler presently addresses fairness between active-state users on times scales of order a second or less, but adding longer-time awareness poses some challenges. The basestation usually 'forgets' users when they transition to idle state, for efficient memory utilization. A stateful basestation would need to retain each user's active and idle context, and handover this information to subsequent cells during mobility events. But when a user leaves the cell while in idle state, the basestation cannot easily determine which cell to forward their context. In contrast the core retains the user state information for long times, and is unaffected by idle-state transitions. Depending upon where in the core the information is retained, context forwarding between regional network elements may need to occur, although such handovers have significantly lower frequency and complexity.

4. The PDF responds to congestion events by pushing congestion management enforcement policies to the PCEF for the subscribers flagged as heavy users on the affected cells, again considering any relevant subscriber information such as currently active premium services.
5. The enforcement policy can take the form of a lower maximum downstream bit rate at the PCEF, or a bearer modification to a lower priority QoS class, which might be reserved for congestion management purposes.
6. When the congestion condition eases, the monitor and analytics functions detect and signal this threshold crossing out-of-congestion to the PDF, which resets (restores to normal) the QoS and/or max bit rate enforcement rules for affected heavy users at the PCEF.

Figure 6.1.2.1-2 (below) illustrates the flow for a Congestion-Aware Fairness use case using ConEx to expose per-flow congestion in real time for a wireless network. In this example congestion detection, marking, and enforcement actions are performed inline, guided by rules set via policy management. It is important to note that ConEx is a work in progress in IETF, and the following is an example of how ConEx could be implemented consistent with current standards-based approaches, but it is not meant to imply that current standards support these ConEx adaptations and extensions.

For any ConEx example, it is assumed that the sending and receiving endpoints are ConEx-enabled. This means that the receiver monitors the ConEx congestion markings and signals this congestion volume back to the sending end via a different (possibly high layer) field or channel, and that the sender likewise monitors this return path indication and returns this information back downstream as ConEx marking within the IP headers. Assumptions about the network support for ConEx-enabled functions are illustrated in the steps of the example below.

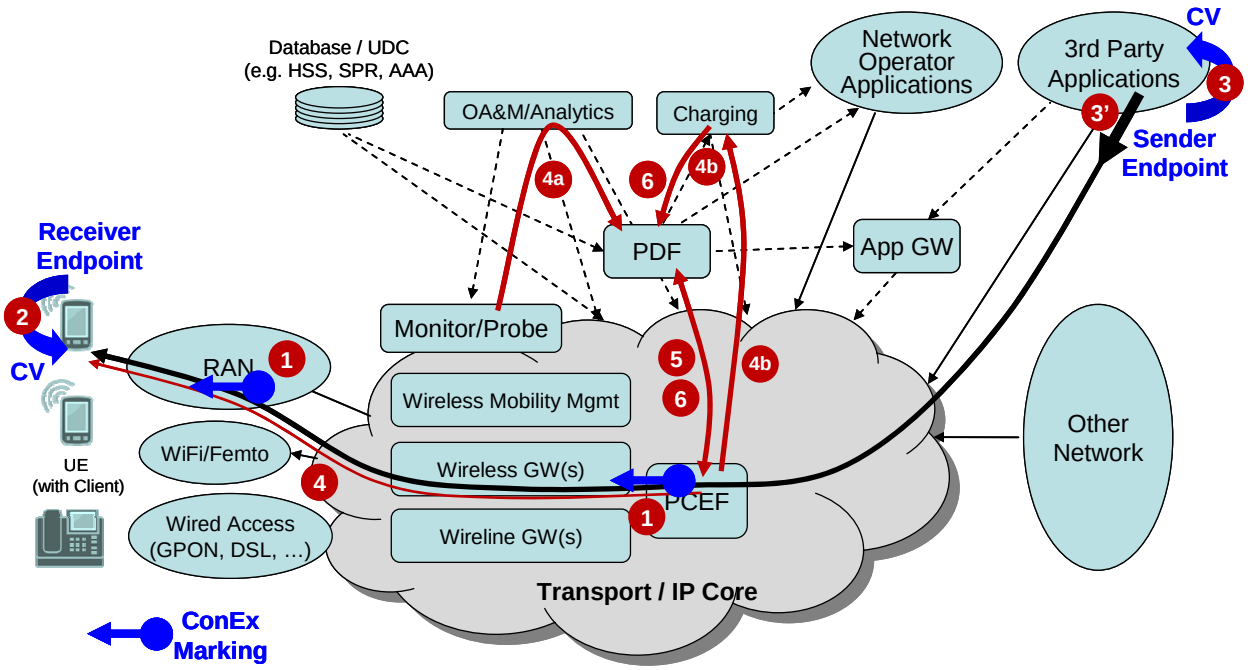

Figure 6.1.2.1-2 ConEx-based Congestion-Aware Fairness

1. At ConEx-enabled enforcement points (e.g., PCEF, RAN, border routers, etc.) along the path of a flow, packets experiencing congestion are ConEx-marked (the definition of "congestion" is beyond the scope of this analysis).

2. The receiver at the terminating endpoint of a flow will signal the accumulated congestion volume in the return path back to the sender endpoint.

3. The congestion volume signaled by the receiver will be reinserted by the sender into the forward path of the flow, thereby indicating to all ConEx-aware nodes in the path the current level of end-end congestion being experienced by that flow.

   Note that, independent of any network enforcement actions, the congestion volume signaled by the receiver back to the sender endpoint allows stronger congestion-aware behavior to be built into the applications themselves. At the sender endpoint, lower priority applications can use this signal to quickly back off the rate of their flow (faster than TCP backoff, similar to the LEDBAT mechanism), while higher priority applications might continue pushing their flow(s), subject to TCP mechanisms for TCP-based traffic (shown as 3′ in the figure). While this mechanism is independent of network enforcement, this enablement of network-friendly application behavior, in combination with incentives via network congestion volume enforcement, is the central concept of the ConEx initiative.

4. At selected points in the network, congestion volume per user will be monitored, accumulated, and thresholded to trigger policy decisions, similar to the non-ConEx Congestion-Aware Fairness example above and/or the usage caps in the User Rate Plans use case. This can be done via monitoring and analytics (4a) or charging systems (4b).

5. The policy decision could include an enforcement changing the inline policing rules for that user, for example reducing the max congestion rate or priority class of a user's traffic after exceeding their monthly congestion volume allotment.
6. Once the billing cycle is reset, or the user upgrades to a higher tier service with a higher congestion volume allotment, a policy decision is triggered that will return the policing rules to what is normal for that subscription tier.

6.1.3 SDO/Gap Analysis

RAN congestion can be inferred in many cases by monitoring standard interfaces. This monitoring function must monitor both user plane and control plane interfaces and must correlate all monitored information to insure that location, throughput, and congestion information is correctly associated with a specific user identity. For example, 3G systems typically expose cell-sector information for radio network control purposes. User plane data can also be monitored where TCP behavior correlated with throughput measurements is used to assess congestion. Taken together, an indication of cell-sector congestion can be inferred. However, in order to accurately measure cell-sector congestion, a standard mechanism to directly expose radio resource utilization outside of the basestation scheduler would be needed.

With LTE, many radio network control functions are integrated into the basestation so that users' cell-sector changes for intra-eNB (evolved Node B) handovers are no longer visible on standard interfaces outside the basestation. This makes inferring congestion based on monitoring and correlating users' TCP behavior and load via only existing standard interfaces less accurate than with 3G systems. This user cell-sector visibility issue could be addressed by extending existing standards to require that intra-cell handovers also be signaled in the same manner as inter-cell handovers, for notification purposes.

6.1.3.1 Area for Standards Development

- Evaluate the feasibility of extending LTE handoff signaling to include intra-eNB handovers.
- Explore the potential for explicit congestion indications using ECN (RFC3168 – Explicit Congestion Notification) or similar mechanisms in the IP header or via standard interfaces from RAN elements. For example, if the basestation can set ECN levels based on air interface resources thresholds, and if the UE supports ECN and can echo this indication back to the sender, monitoring capabilities in the network can better determine the RAN congestion state of user flows.
- The determination of heavy users using the newly defined 3GPP Sy interface between the OCS and the PCRF may require Sy enhancements to perform this function.

For backhaul and/or wireline access, existing mechanisms and protocols that may be incorporated into Congestion-Aware Fairness solutions include the following:

- IP Flow Information Export (IPFIX) Protocol, specified in RFC5101 and related documents, may be useful for per-flow volume monitoring.
- OA&M performance monitoring mechanisms, including Institute of Electrical and Electronics Engineers (IEEE) 802.1ag and International Telecommunications Union-Telecommunication (ITU-T) Recommendation Y.1731, provide delay and packet loss metrics that may indirectly indicate congestion events.

- RFC3168 describes ECN that may provide both direct indication of congestion events and indication of the flows contributing to those events.

Note that some of the existing mechanisms listed above are protocol-specific. In addition, the current usage of some of the above mechanisms by Element Management Systems (EMS)/Network Management Systems (NMS) may not be oriented towards real-time monitoring and time stamping of events as envisioned for congestion-aware fairness mechanisms. If different EMS/NMS and/or monitoring systems are used to detect congestion for different parts of a network (e.g., RAN and backhaul), then interfaces between these systems may need to be standardized in order for congestion to be correlated between network segments and indicated to the PDF in an aggregated and consistent manner.

For the purposes of congestion-aware fairness, the above EMS/NMS mechanisms typically do not identify which users are impacted by the congestion event. ECN mechanisms mark the congestion in the IP headers and thereby allow network monitoring capabilities to better correlate the user to a congestion event in near real time.

Regardless of how congestion and subscriber usage information is detected, a standard way of providing these triggers to the PDF is needed.

6.1.3.2 Future Area for Standards Development

The ConEx working group within IETF is working to develop a mechanism by which senders inform the network about the congestion encountered by previous packets on the same flow. Primary work items include:

- An Informational document containing an abstract description of the congestion exposure mechanism that is independent of specific transport protocols and congestion information encoding techniques needed for different IP protocol versions. (July 2011)
- An Experimental specification of an Ipv6 packet structure that encapsulates ConEx information, defining a packet format and an interpretation. (November 2011)
- An Experimental specification of a modification to TCP, for the timely transport of congestion information from the destination to the sender. (November 2011)

ConEx work currently underway includes: draft-ietf-conex-abstract-mech-01, "Congestion Exposure (ConEx) Concepts and Abstract Mechanism" and draft-ietf-conex-concepts-uses-01, "ConEx Concepts and Use Cases." A separate Internet draft, draft-kutscher-conex-mobile-00, "Mobile Communication Congestion Exposure Scenario," explores issues specific to implementation of ConEx in mobile networks. The white paper "Internet: Fairer is Faster" (FairerFasterWP.pdf) provides a useful tutorial.

The PCC architecture will need to be extended to include congestion volume-based (as opposed to simple volume) accounting and enforcement policy rules, as well as ConEx-aware policing functions in user plane enforcement point elements. It is not expected that major structural changes to the PCC architecture would be required, but rather that these extensions would be additional cases and parameters added to existing functions and interfaces. It is also likely that more enforcement points will be needed in a ConEx-enabled network.

6.2 Subscriber-/Application-Aware Network Optimization

6.2.1 Service/Business Needs

In the analysis of the Subscriber-/Application-Aware network optimization scenario, the following service needs and capabilities were considered.

SERVICE CAPABILITY: The Network shall have the ability to apply policies to specific data flows to optimize network performance for subscribers and applications.

SERVICE CAPABILITY: The network shall provide a means to allow subscribers to request network optimization.

SERVICE CAPABILITY: The network shall provide a means to allow Application Provider to request optimization.

SERVICE CAPABILITY: The network shall acknowledge the request indicating the success or failure of the optimization request.

SERVICE CAPABILITY: The network shall be able to detect that a flow exists for a specific subscriber and/or application.

SERVICE CAPABILITY: Once a flow is detected, the network shall be able to determine:

- Flow description (at least the 5-tuple);
- Start and stop events for enabling/disabling the optimization when applicable;
- Application Service Provider Identity if Application Provider-specific policy and charging rules are to be applied to the service; and
- Subscriber Identity if subscriber-specific policy and charging rules are to be applied to the service.

SERVICE CAPABILITY: The network shall be able to make a policy-based decision for optimization on the specified flow, as follows:

- The network shall arbitrate between the specific optimization policy and other policies in force for the specific flow; and
- The network decision will include both policy and charging rules.

SERVICE CAPABILITY: The network shall be able to enforce the policy decision to enable the optimization. Policy enforcement may include:

- Adjusting QoS priority level;
- Adjusting policing parameters for traffic;
- Traffic steering/routing rules; and
- Traffic/flow optimization (e.g., video transrating, transcoding, caching, or TCP optimization).

6.2.2 Implementation Options

There are many different possible use cases and implementation approaches associated with the Subscriber-/Application-Aware Optimization scenario. In the example outlined below, a wireless subscriber has opted into a network service that provides video or other application-specific optimization when the user is in limited bandwidth conditions. These conditions could

be caused by wireless air interface signal limitations and/or congestion in some part of the network.

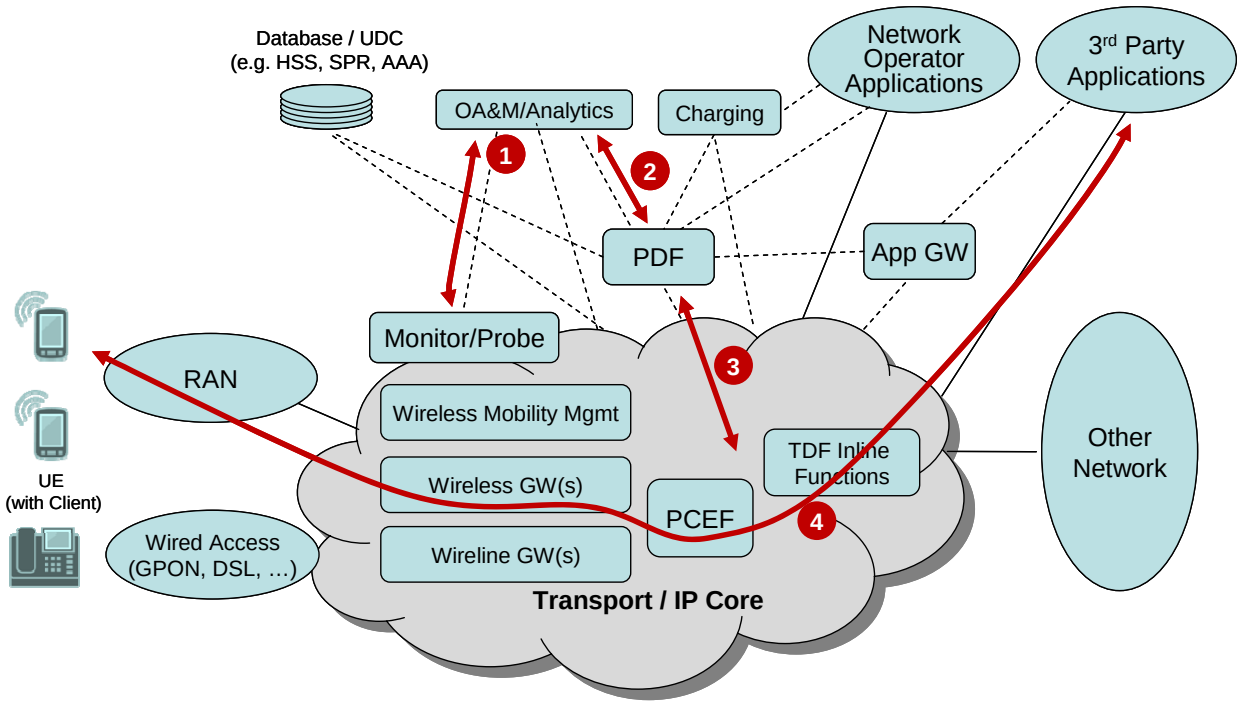

Figure 6.2.2-1 Subscriber Based Content Optimization

1. The Analytics function, in cooperation with the Monitor/Probe functions, determines that a subscriber is experiencing congestion.
2. The Analytics function recommends to the PDF a new subscriber policy to steer subscriber flows through one or more application specific traffic optimization functions. The PDF will verify network and subscriber policy to ensure consistency (i.e. that the subscriber has opted into the service).
3. PDF will communicate a steering policy to the network to route subscriber flows to one or more optimization functions. The PDF may also transfer rules to the TDF specific to the specific optimization required. For example, an application-specific user priority (or optimization index) could be transparently passed to the TDF to enable subscriber specific levels of optimization.
4. Subscriber flows are routed through optimization functions.

Note that the steering policy can be removed through the same sequence of steps once the Analytics function determines that the subscriber no longer needs specific traffic optimization functions.

Other examples of applications and subscriber-aware optimization, not shown in detail here, include:

- A user opts into a data plan that offers special treatment for specific applications (e.g., traffic receiving lower priority in exchange for reduced charging);
- A user subscribes to a service that routes all traffic for specific users to a Parental Controls service;
- Transcoding: a user transfers video session from HD TV to a mobile device; policy introduces transcoding to match video quality to the capabilities of the device; and
- The use of IETF ALTO (Application Level Traffic Optimization) to optimize the delivery of P2P traffic.

6.2.3 SDO/Gap Analysis

For wireless networks, 3GPP Releases 8 and 9 already support major aspects of application and user differentiation. For example, the Rx protocol between the PCRF (PDF) and various application functions supports media description for each component that comprises a multimedia session such as voice, video and data. Each media component includes priority relative to other components. The Rx interface also includes an application identifier to assist the PCRF in making policy decisions appropriate for the application. The subscriber's priority is represented in the reservation-priority Information Element (IE). In addition, the PCRF derives the Allocation Retention Priority (ARP) IE and the QCI based on the Rx and Gx (PCRF – packet data network gateway (PDN-GW) PCEF interface) Ies, the subscriber's profile, other information that may be available to the PCRF. The PCC rule for each service data flow of a multimedia session includes the ARP and QCI. The priority component of the ARP unambiguously represents the user's priority and can vary depending on the service. For instance, the ARP for voice can be higher than the ARP for video for the same user.

Although application and subscriber priority indicators are available, the inline optimization function can often make use of an application-specific user priority to allow varying degrees of optimization for each application. This application-specific priority could be associated with each application identifier, which could be passed transparently from the SPR to the inline function. Further study is required to determine if existing information elements can be used for this function.

Other aspects of user differentiation include APN/destination network Aggregate Maximum Bit rate (AMBR) and UE AMBR for all UE connections. Both APN and UE AMBR apply to non-Guaranteed Bit Rate (GBR) services.

3GPP Release 11 will include the TDF.

In addition, applications can subscribe to events such as loss of signaling RAN-type change, unsuccessful allocation of resources, loss/recovery of bearer, or out-of-credit-indication.

For optimizations involving over-the-Internet requests from 3rd party applications or subscribers, these requests can be carried over network interfaces between the two parties, network APIs, or a combination of the two. While API standards exist for more mainstream use cases such as QoS-ensured connections, there is a virtually unlimited range of possible subscriber- and application-aware variants, many of which would require new or extended network API definitions.

6.3 Network-Aware Scheduling of Content

6.3.1 Service/Business Needs

In the analysis of the Network-Aware Scheduling of Content network optimization scenario, the following service needs and capabilities were considered.

SERVICE CAPABILITY: The solution needs to be aware of network load and/or throughput capabilities.

- Such awareness could be obtained via measurement of network throughput, sensing of congestion, and/or advance signaling.
- Predictive scheduling (e.g., time of day) is by itself is likely insufficient, as different parts of the wireline or wireless access network (e.g., cells) are known to have very different busy hours. However, there may be some benefit through simple schemes such as night-time delivery, possibly augmented by awareness of individual user patterns.
- Additional future improvements could include signaling to the UE of end-to-end network load and/or available throughput.

SERVICE CAPABILITY: The network shall be able to identify if a specific user will contribute to or be impacted by congestion identified above.

SERVICE CAPABILITY: The solution needs to support situations where the source content is served by the NSP network or from external an application server.

6.3.2 Implementation Options

Option 1: UE and Application Scheduling Portal Driven

One approach is to have a software agent/client or application on the UE that would support time-shifting.

- Such a client is likely the simplest way (i.e., no network interworking required) to address time shifting in environments that span multiple technologies, multiple accesses, and/or multiple service providers.
- The agent would be aware of when different access technologies are available (e.g., Wi-Fi). (See section 6.6 for further detail on multi-access selection.)
- With future standards changes, additional improvements are possible. For example, the RAN could also signal end-to-end network load and/or available throughput, to further assist the UE agent's performance.

The UE client/application could access a network-based application server to manage time shifting of content using a flexible rate plan that allowed a subscriber to request a lower charging rate for traffic exchanged during periods of low congestion. In this approach, the application, either from a UE client or directly from the application server in the network, would request a "low congestion rate" from the network to lower the cost of the data transfer. This request could be triggered by the need to download a large file. The application could defer the download for some specified period of time if a lower rate was refused. Alternatively, the application could choose to download the material at the standard charging rate at any time. Examples of potentially deferrable downloads include weekly anti-virus updates, application software updates and side-loadable content requests.

Approach 1:

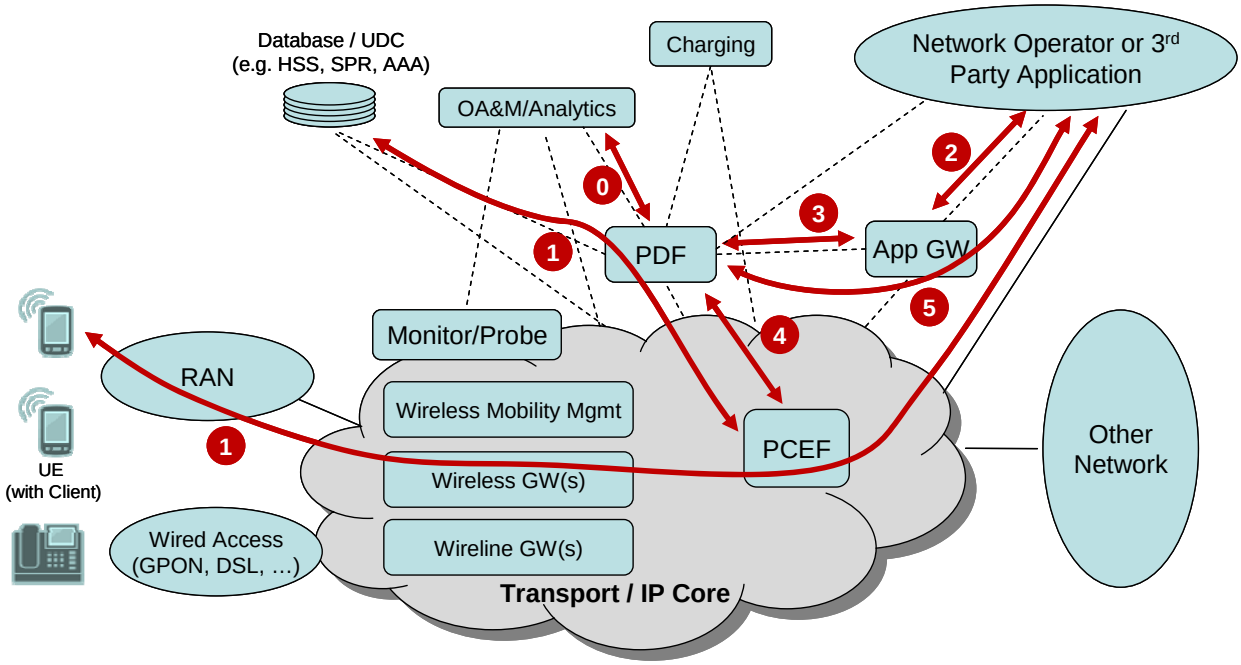

Figure 6.3.2.1-1 Rate Plan Driven Deferred Scheduling

0. The PDF receives network status information from OA&M/Analytics on a periodic basis. Alternatively, the PDF polls the OA&M/Analytics periodically for load status of the network.

1. The UE attaches to the network and the PDF obtains the UE's profile from the SPR. The subscriber policy allows the flexible rate plan approach as outlined in this use case. An application is triggered to download content. The specific trigger can be from a network application server directly or from a client resident in the UE.

2. If the Application is untrusted (3rd party), the application requests a lower charging rate via the AppGW. The AppGW provides the security functions, authorization, charging and service level agreement (SLA) enforcement for all applications per OMA and/or 3GPP specifications. Alternatively, a trusted application could communicate directly with the PDF as in step 3.

3. The AppGW or trusted application sends the authorization request to the PDF. The PDF examines the request and depending on the load status of the network, user profile and other variables, the PDF: a) accepts the request; b) proposes a re-try interval. The re-try interval can vary depending on the network load status and other policies including time-of-day.

4. If the request for a different charging rate is approved, the PDF can specify new charging rules in the PCEF. The PDF may revert to the default charging rules after a specific period of time. The Application/UE/Client would need to re-authorize prior to that time to continue to receive the lower charging rate.

5. Application/UE/Client is informed of the decision. If the request is denied, the UE/Client could be provided the re-try interval for re-request. If the request is honored, the period of time for which the new charging rules are enforced would also be conveyed.

Further study is needed to determine the enhancements to network APIs (OSA/Parlay-x/RESTful) to allow requests for lower charging rate access, and for negative responses to indicate a re-try interval. Enforcement of the re-try interval by the AppGW remembering that the request from the 3rd part application is a re-try and ensuring that the re-try is not attempted before the re-try period has expired should be considered for implementation.

Approach 2: ECN-Based UE-Driven

In this option, ECN code points in the IP header are used to signal congestion to the wireless UE, allowing the UE to respond appropriately.

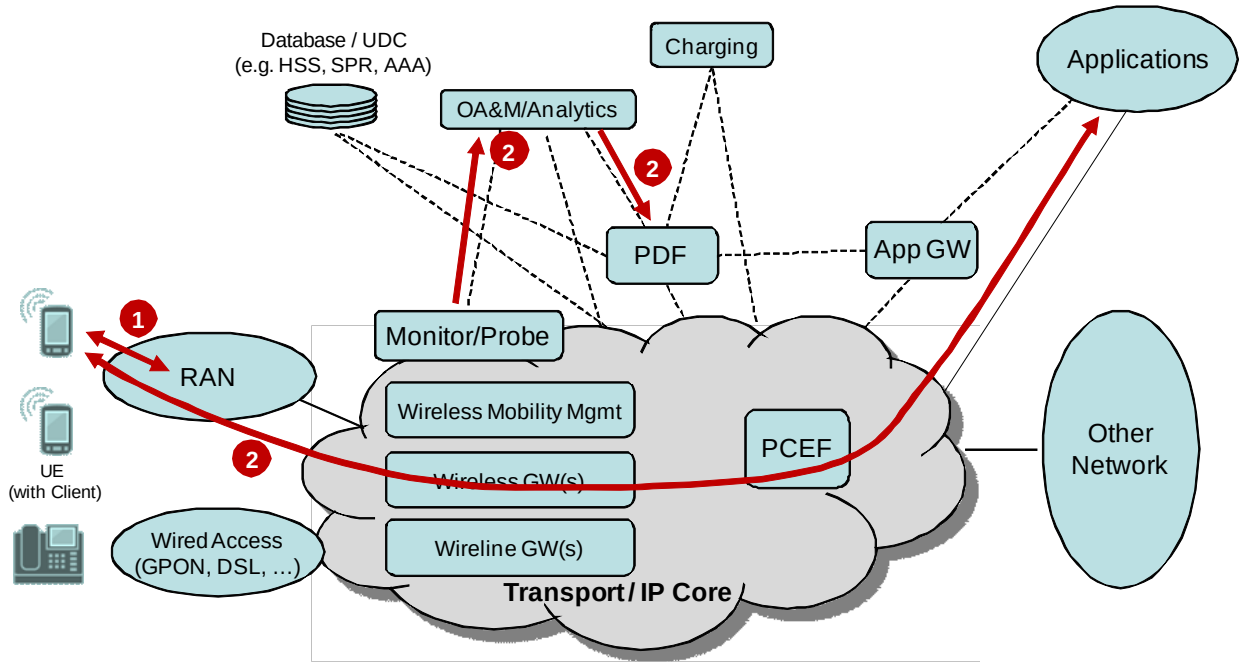

Figure 6.3.2.1-2 ECN-Based UE-Driven Deferred Scheduling

The source of the ECN notification can be the basestation when congestion exists over the air, or can be other areas of the network such as the backhaul/wired access network or the core network elements. The ECN notification applies to both UDP and TCP traffic. Once the UE has received a notification of congestion, it can (via the UE operating system) interact with the application domain to defer data transfers that are deferrable or further rate limit non-critical traffic.

In addition, the UE could use ECN-Echo and CWR flags in the TCP header to signal the presence of congestion back to the TCP sender/server in the network. This information could be intercepted by the network to aid in the determination of cell related congestion for other Network Optimization use cases. This option is highlighted in step 2 in the figure above.

6.3.3 SDO/Gap Analysis

It is clearly desirable for the network to support PCC functionality such as QoS and associated policies. This is not rigorously required, but clearly provides additional complementary functionality and benefit via policy decisions and enforcement.

The opportunistic scheduling of a service will likely be controlled at the service layer. The application of transport layer QoS and the associated charging data collection performed by the 3GPP PCC framework will occur as for any other service. At the billing stage, the charging data collected (e.g., the number of bytes consumed for the service delivery and the transport QoS used for it) will be associated with the service layer information (e.g., download of a newspaper based on opportunistic scheduling). Any billing impact because of differences in service offering between a real-time delivery and a time-delayed/opportunistic delivery are backend IT aspects that do not have standards impacts.

The PCC-related deferred scheduling scenarios benefit from a re-try interval parameter to better manage repeated requests for deferred scheduling. Areas for standards development include:

- A new re-try interval parameter on the Rx reference point
- Potential enhancements to the Application Gateway API (OSA/Parlay-x/RESTful) to indicate a re-try interval. Enforcement of the re-try interval by having the AppGW "remember" that the request from the 3rd part application is a "re-try" and ensure that the re-try is not attempted before the re-try period has expired is an implementation consideration.

Approach 2 requires support for ECN in both the basestation and the UE. For example, the basestation could set ECN congestion based on a rolling average of air interface radio resource utilization. The time and threshold parameters used to trigger congestion via ECN could be operator provisionable parameters allowing for some control and management of air interface resources. Any flexibility designed into the use of ECN, however, needs to be consistent with IETF guidance. For TCP traffic, as provided in RFC3168, the congestion control response would occur in the transport layer; for UDP traffic, the response would occur in the application layer.

6.4 User Rate Plans

6.4.1 Service/Business Needs

In the analysis of the User Rate Plans network optimization scenarios outlined in Section 4.4, the following service needs and capabilities were considered.

SERVICE CAPABILITITY: The network shall have the ability to factor a subscriber's service tier into service parameters (maximum uplink and downlink bite rates, default QoS priority, charging rules, etc.) of that user's access connection and usage thresholds (i.e. monthly usage cap).

SERVICE CAPABILITITY: The network shall have the ability to detect tier-specific usage volume thresholds, and dynamically change the service parameters for that user while they are over-threshold. It should also have the ability to alert the user (via SMS, or http redirect) when they are at or approaching that threshold, providing options to increase their usage threshold and/or service tier.

SERVICE CAPABILITITY: In ConEx scenarios, the network shall be able to measure congestion volume per user, and be able to threshold and apply tier-specific enforcements

similar to the (usage volume-based) service capabilities listed above, or in the form of setting ConEx-enabled policing policies at enforcements points (e.g., to perform enforcements selectively to ConEx-marked traffic). For wireless networks, congestion volume in the RAN may need to be calculated in a special way to accurately reflect a user's impact on congesting radio resources.

SERVICE CAPABILITITY: Users shall have visibility to the measured parameters (i.e. usage volume or congestion volume) used to police and/or charge for that user's consumption of resources on the network.

6.4.2 Implementation Options

In the following example implementation, different subscription tiers are differentiated by the amount of data volume that is included, with rate-limiting enforced beyond that level, and by the inclusion of zero-rated access to content from a specific online content aggregator for a special bundled premium tier.

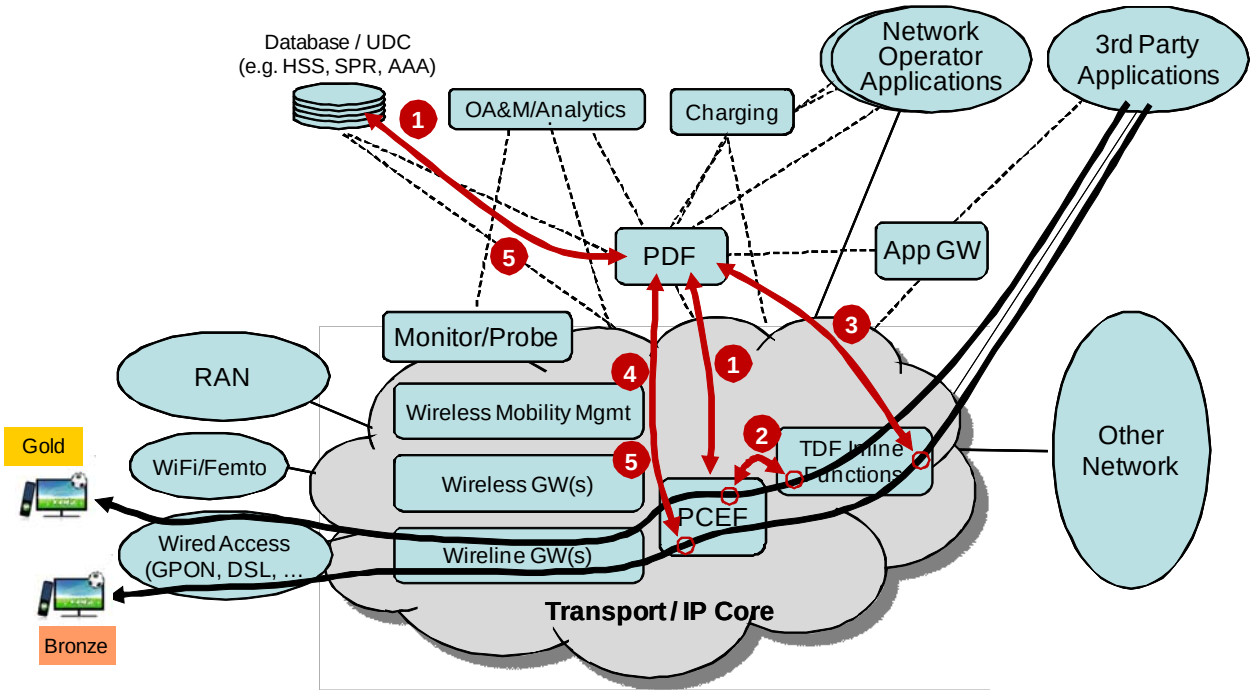

Figure 6.4.2.1-1 User Rate Plans

1. Two wireline broadband subscribers sit down for an evening of online HD sports programming, one has a basic "Bronze" subscription and the other has a premium "Gold" package that includes unlimited viewing for a specific set of online HD sports channels. When the users' broadband modems last powered up and attached to the network, policy rules based on their subscriber profile were signaled by the PDF to the PCEF in the Broadband Network Gateway (BNG) serving their location.

2. When the Gold user starts streaming bundled HD programming, the TDF detects that the flow is for an included service, and if the TDF is part of the PCEF the detected flow is not counted toward the user's usage cap (per the pre-loaded policy rules). This may be done with simple IP address filtering, since the services are sponsored and well known.

If the TDF is a standalone element separate from the PCEF, a mechanism is required to allow the detected flow count to be subtracted from usage information provided by the PCEF. This can be implemented by having the standalone TDF use the sponsored data connectivity feature as documented in 3GPP TS 23.203. In this case, the TDF would provide an Rx interface to the PDF/PCRF to represent the detected flow as being sponsored.

3. As the Bronze user streams the same content, the pre-loaded policy rules count these flows normally. When the monthly usage cap is exceeded (due to significant consumption of such HD content), the PDF is alerted.
4. The PDF responds by checking the subscriber's profile and making an enforcement decision—in this case, to (a) modify the enforcement rules in the PCEF to significantly limit that user's maximum downstream bit rate and (b) redirect the user's session to a portal where the subscriber can upgrade to a higher subscription tier providing a larger monthly usage cap and/or premium bundled online HD sports content.
5. If the Bronze user upgrades to the Gold tier, then the user's subscriber profile will change, the PDF will update the user's policy rules in the PCEF, and the user can resume streaming with high quality.

6.4.3 SDO/Gap Analysis

The use of policy-controlled rate plans is a relatively mature topic area, especially for wireless networks. As a result, no significant additional areas for standards development (beyond ongoing converged policy control architecture) were identified for this use case area. Annex N (Informative) of 3GPP TS 23.203 provides a good overview of the sponsored data connectivity feature, which can be used for zero-rating use cases.

6.5 Reasonable Network Protection & Management

6.5.1 Service/Business Requirements

In the analysis of the Reasonable Network Protection & Management network optimization scenario, the following requirements were considered.

SERVICE CAPABILITY: The Network shall have the ability to detect harmful traffic (e.g., DoS attacks, malware, or rogue user behavior) and illegal traffic. The detection mechanism can be network-based or device -/client-based where the client signals to the network upon detection of such traffic. In this case, the client should be trusted by the network to provide this detection mechanism. Harmful traffic may be targeted toward users or network elements such as network servers, DNS servers or other network signaling and routing elements.

SERVICE CAPABILITY: Upon receiving an indication of harmful/illegal traffic from either a network-based detection capability or a trusted client based detection capability, the network must have the ability to quarantine the impacted flow and/or throttle the flow based on network policy.

- When possible, the end-user sourcing the detected traffic flows should be notified of such flows and of any resulting network actions.

6.5.2 Implementation Options

In the following approach, user traffic is routed to a TDF inline function to screen for harmful/illegal traffic. If this type of traffic is detected, the PDF in the network is invoked to

apply policies to quarantine and/or throttle the impacted flows until the issue has been resolved.

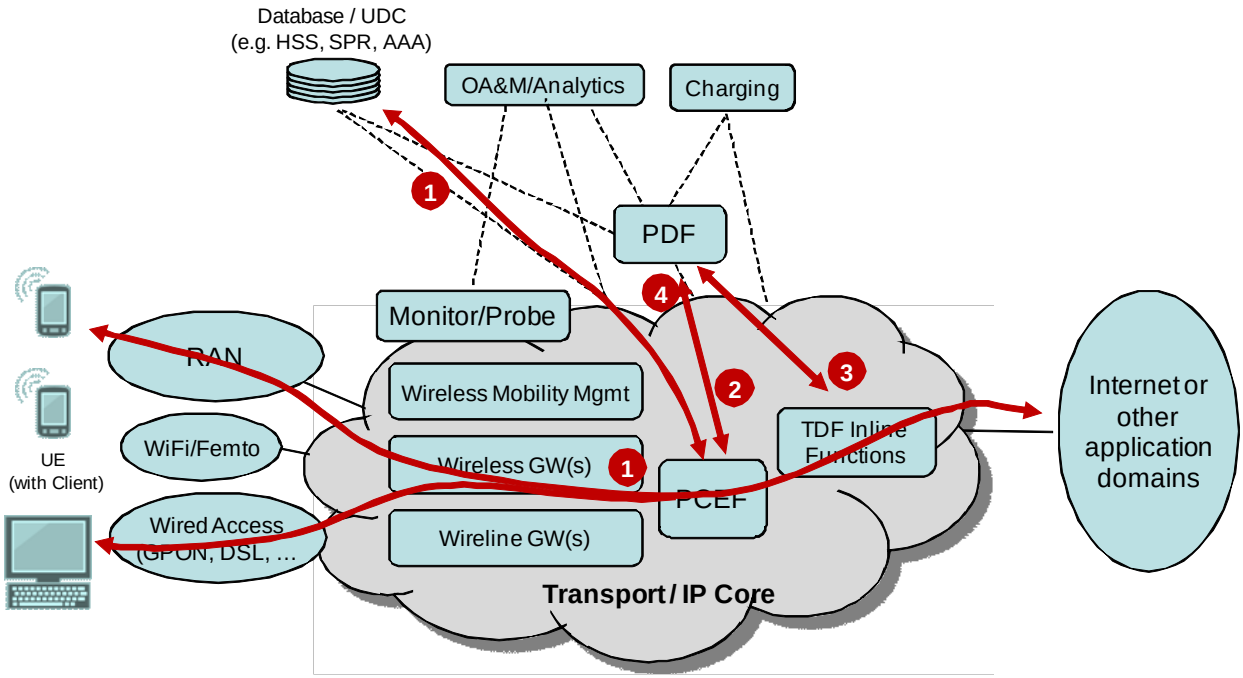

Figure 6.5.2.1-1 Reasonable Network Protection & Management

1. The UE attaches to the network and the PDF obtains the UE's profile from the SPR. The subscriber policy allows users to opt-in to the network protection service if necessary. Alternatively, some or all of the network protection capabilities may be enabled to provide "always on" protection.
2. Based on PDF decision, a policy is forwarded to the PCEF to steer traffic through an inline TDF function used to detect harmful/illegal traffic flows.
3. If harmful/illegal traffic is detected, the PDF is notified.
4. PDF will forward a policy to the PCEF to route the affected flow to a quarantine server (not shown in the figure). If possible, the subscriber sourcing the affected flows should be notified of the detection and any policy enforcement. This could be done by the TDF Inline function. Future study is needed to determine under what conditions the PDF will reset the steering policy to allow traffic to flow to the appropriate application domain.

Monitor/analytics-based detection (not shown above) is another option for detecting wireless-specific threats, such as signaling attacks.

6.5.3 SDO/Gap Analysis

No standards impact has been identified to implement the needed functionality. However, the application-specific priority parameter (optimization index) discussed in Section 6.2.3 (Subscriber-/Application-Aware Network Optimization) could also be applicable in this case.

For example, the application-specific priority could specify different levels of malware detection (e.g., varying the aggressiveness of the detection capabilities).

6.6 Load- and Policy-Aware Multi-RAN Selection

6.6.1 Service/Business Needs

In the analysis of the Load- and Policy-Aware Multi-RAN Selection network optimization scenario, the following service needs and capabilities were considered.

SERVICE CAPABILITY: The access network must be able to monitor, on an appropriate granularity, its loading level and/or available throughput (e.g., for cell-sector, backhaul, and associated radio network controller (RNC)/mobility management entity (MME), for downlink and uplink).

SERVICE CAPABILITY: To facilitate inter-technology access selection or load balancing, the cellular basestation or access point would broadcast or signal a coarse-grained load (or available throughput) indicator to UEs, for example on an existing overhead channel. The time scale for this broadcast (and the underlying load/throughput metric) can also be coarse, e.g., every few seconds. The UE (together with an embedded client or application) can then make its attachment/connection choice, based upon service provider policies.

SERVICE CAPABILITY: The solution needs to support multiple access technologies (e.g, cellular and Wi-Fi access), for purposes of network selection as well as for application routing flows.

SERVICE CAPABILITY: The solution needs to support network configurations where multiple service providers are involved (e.g., macrocell NSP, pico/femto-cell NSP, roaming NSP, wireline NSP supporting a femtocell and/or Wi-Fi access points, etc.).

6.6.2 Implementation Options

If similar loading (or available throughput) information is not made available for certain technologies (e.g., Wi-Fi), the UE could potentially sense Wi-Fi access point availability (utilization, interference levels, etc.) to make more informed access selection decisions.

The concepts discussed here can in principle be applied to both uplink and downlink, but actual implementations initially might only address one direction (e.g., today the downlink is typically more heavily congested).

The advent of smartphones has complicated the determination of access congestion. Most of these devices will have the capability to access more than one radio access technology (e.g., 2G, 3G, or Wi-Fi). Within standards, the capability exists to steer traffic on to these different networks. Within 3GPP, a function has been defined called the Access Network Discovery and Selection Function (ANDSF). The ANDSF allows the operator to download Inter-System Mobility Policies (ISMP) and Inter-System Routing Policies (ISRP) to a device. These guide the device in deciding which accesses to use when and for which flows. In general, the functionality of the ANDSF is continually being enhanced (e.g., 3GPP Operator Policies for IP Interface Selection (OPIIS) and Data Identification in ANDSF (DIDA)) to allow more flexibility in how to select routing. There is currently no connection between the PDF and the ANDSF, so these routing decisions are currently taken independently and may interact with congestion policies implemented at the PDF. ANDSF is defined in 3GPP 23.402.

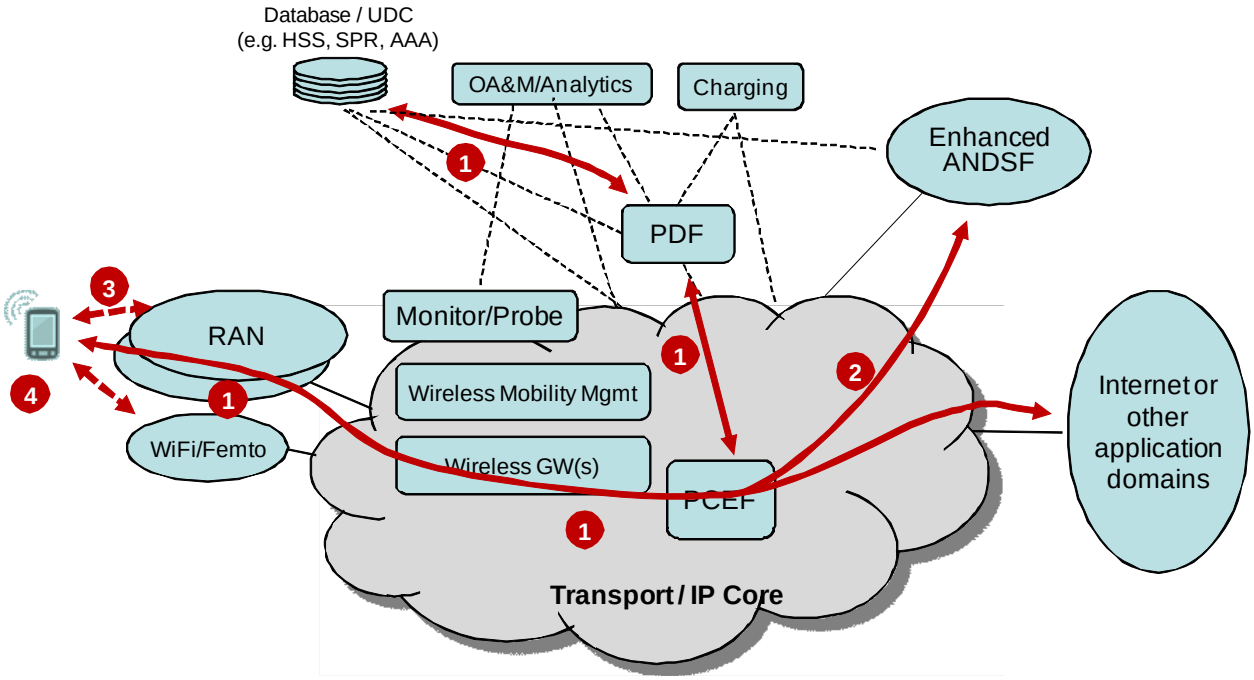

Figure 6.6.2.1-1 Load- and Policy-Aware Multi-RAN Selection

1. UE attaches to the network. As part of the existing attachment procedure, the PDF (PCRF) is contacted to install appropriate policies and charging rules. The SPR (using UDC) may be queried in the process to get subscriber specific policies.
2. Once attached, the ANDSF can provide updated access network selection policies. The ANDSF will also use UDC to retrieve subscriber policies and should have access via UDC to the same subscriber policies available to the PDF to insure that device policies are consistent with network policies for this specific device and attachment instance.
3. RAN (and potentially other non-licensed RANs) broadcasts a coarse-grained measure of bandwidth availability (based on network load) over the air interface.
4. The UE, as required based on signal conditions, can use ANDSF rules along with RAN availability measures to make a decision on which access network(s) to attach to and/or direct application flows to.

An alternate implementation (not shown) for steering/directing application routing flows would be to use network APIs to expose congestion detected in the network (similar to use cases in Sections 6.1, 6.2, and 6.3). This does not require basestation changes, but does not address initial network attachment.

6.6.3 SDO/Gap Analysis

Intelligent access decisions between cellular and Wi-Fi would require standardization of network mechanisms to communicate load (or available throughput) indicators to the device or other core network elements. There are also standards implications for the devices, specifically how they receive these load indicators and how they might act upon them (e.g., applying any policies regarding access technology selection).

For Wi-Fi networks, amendment k to IEEE 802.11 defines an air interface load indication. Specifically, the channel load request/report pair returns the channel utilization measurement

as observed by the measuring access point. However, since the air interface is often faster than the backhaul link in common Wi-Fi networks, an indication of the air interface may not be useful in assessing how good the throughput is going to be. Wi-Fi Alliance Hotspot 2.0 has been looking into advertising to the UE the backhaul capacity. Although statically configured backhaul link speed could be used in this regard, it would be more useful to advertise some measure of estimated throughput based on recent link load information.

To complement Wi-Fi based load indications, 3GPP could have the RAN (cell or cell-sector) advertize some measure of its load (or alternately, its available throughput) averaged over some coarse time scale (e.g., 10 seconds – 1 minute) which a UE agent or client could utilize for optimal access network selection. Such a scheme reduces the implied control channel impact while providing information on a more relevant time scale. These enhancements would require enhancements to 3GPP RAN standards.

Similar to the Congestion-Aware Fairness case, accurately detecting congestion levels may require correlating congestion information from both RAN and backhaul. If different EMS/NMS and/or monitoring systems are used to detect congestion for different parts of a network (e.g., RAN and backhaul), then interfaces between these systems may need to be standardized in order for congestion to be correlated between network segments and indicated to the PDF in an aggregated and consistent manner.

It is expected that many devices will be able to dynamically utilize both 3G/4G accesses and Wi-Fi depending on factors such as availability and utilization levels on the accesses, application requirements, operator policies, and user requirements. These rules will be instantiated in the ANDSF. While these rules are relatively static, the actions arising from the rules may be dynamic. It is recommended that ANDSF and PCC derive their data from a common UDC (e.g., data used to synthesize policies such as application identifiers).

ANDSF provides policies that govern access selection (3G/4G vs Wi-Fi) based on various criteria such as:

- Type of access
- Access network identifier
- Location
- Time of day
- IP Flow (5-tuple)
- Application characteristics (addressed in 3GPP Release 11 specifications)

It is proposed to enhance the list of policy criteria to include congestion information on the accesses (e.g., use Wi-Fi access if RAN load is above 50%).

Congestion exposure using the network API would require additional API development in OMA or other SDOs.

6.7 Optimizing Use of Wireless Non-Bearer Resources

6.7.1 Service/Business Needs

Networks have been optimized to meet the signaling loads associated with human initiated communications which is mainly voice and simple data applications. However, smartphone

applications and machine type communications often impose different usage profiles from those associated with human communications, and those are the main reasons to trigger the need for non-bearer resource optimization. Therefore, the service needs and capabilities appropriate for this scenario will be more based on those two deployments.

SERVICE CAPABILITIY: The network shall provide a mechanism to restrict access to signaling resources from low priority MTC devices when the network is congested.

SERVICE CAPABILITIY: Signaling shall be optimized to support large numbers of MTC devices.

SERVICE CAPABILITIY: The system shall be optimized to handle MTC groups. It shall be possible to associate one MTC device to more than one MTC group. Also, the network shall provide a mechanism to send a broadcast message to members of an MTC group.

SERVICE CAPABILITIY: Congestion controls can be applied to groups of MTC devices.

SERVICE CAPABILITIY: Restriction of access for MTC device is triggered when the network is congested.

SERVICE CAPABILITIY: Signaling may be optimized for devices that have low or no mobility, such as implementing a management procedure to reduce mobility for these kinds of devices.

SERVICE CAPABILITIY: Signaling optimizations may be provided to cater to time-scheduled MTC devices or smartphone applications.

SERVICE CAPABILITIY: Signaling shall be optimized for those MTC devices that do not support (or rarely support) mobile-terminated communications (mobile-originated only).

SERVICE CAPABILITIY: Signaling shall be optimized for those MTC devices that communicate infrequently.

SERVICE CAPABILITIY: MTC devices that are marked as time-tolerant may not even be permitted to attach to the network during congestion.

SERVICE CAPABILITIY: Signaling may be optimized for data-only devices.

SERVICE CAPABILITIY: Overload controls may be provided to prevent MTC devices or smartphone applications from causing signaling overload of the network (either from periodic events or exceptional events such as a mass reattach caused by a network outage).

SERVICE CAPABILITIY: Signaling may be optimized for devices or smartphone applications doing small data transmissions. These include optimizations for status messages and keep-alive messages.

SERVICE CAPABILITIY: Congestion mechanisms shall exist to prevent overload of nodes and allow graceful degradation of traffic during overload conditions.

SERVICE CAPABILITIY: When one or more network nodes experience signaling congestion, the network should minimize or eliminate the impact on the other devices or applications whose signaling transactions are not through those congested nodes.

SERVICE CAPABILITIY: The networks and devices should provide application developers with the APIs and tools necessary to make efficient use of access and network resources. Application developers should make use of these tools.

6.7.2 Implementation Options

These issues apply primarily to wireless accesses and their core network since these have the highest signaling load with very limited network resource. Some of the behavior is associated with flags set by the device and communicated over the access network. This allows the access network to differentiate and optimize for different classes of devices.

One mechanism defined so far is extended access barring, which allows the operator to bar access to groups of devices. This mechanism is independent of traditional access barring classes implemented in 3GPP systems.

Signaling congestion can be triggered by device or application behavior, such as aggressive access or polling. There are two types of solutions. The network can passively throttle down the signaling traffic by restricting or stopping access to the device or the network. Alternatively, the device or application can coordinate with some open API to reduce or schedule the signaling traffic. This option is demonstrated by the following two potential implementation options:

Option 1: UE heartbeat for keeping firewall ports open

Many applications use heartbeat mechanisms where a UE periodically sends a short message to the application server to indicate it is online. But the current heart beat mechanism is purely transparent to the network as normal data plane traffic. As such, in a mobile network, the UE will constantly be switched between idle and active state in order to send those small packets, causing unnecessary network signaling traffic and reduce UE battery life. In this example, UE application will make use of capabilities in the messaging API which allow the UE to intelligently group the heartbeat messages to efficiently use of the radio interface.

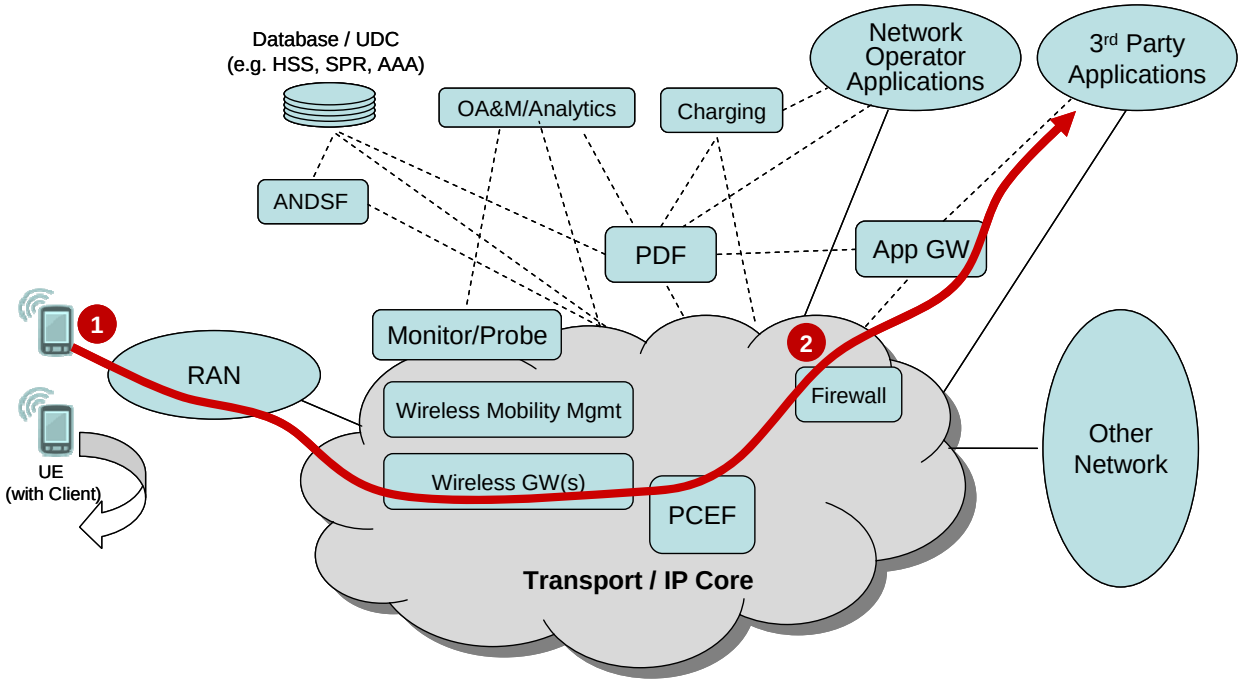

Figure 6.7.2-1: UE heartbeat for keeping firewall ports open

1. The application clients embedded in the UE that are ready to send a heartbeat to the server will first deliver the heartbeat message to the application support layer or middleware in the device. The application support layer or middleware function will batch the heartbeat messages so that the radio interface and battery consumption is minimized.
2. The heartbeat message will keep the firewall pinhole open and enable messages to be exchanged between the application in the UE and the application server.

Option 2: network-API-based gateway for checking UE status and out-of-firewall pushing

In the following example of a potential implementation option, the smartphone application interacts with the network using a network API to get a user's connection status instead of running its own IP-based heartbeat. Additionally, the network can use information from this API call to give the application server the user's connection quality information, allowing the application server to adjust accordingly. For some short and frequent pushing or paging service which may be out-of-firewall, the application gives the push/paging notification to the network, and the network can push the data while the user is in an active state or can use a more signaling-efficient way to communicate with the device. The goal is to avoid a situation where the device goes into idle mode and then must re-establish the connection to pass a small amount of data.

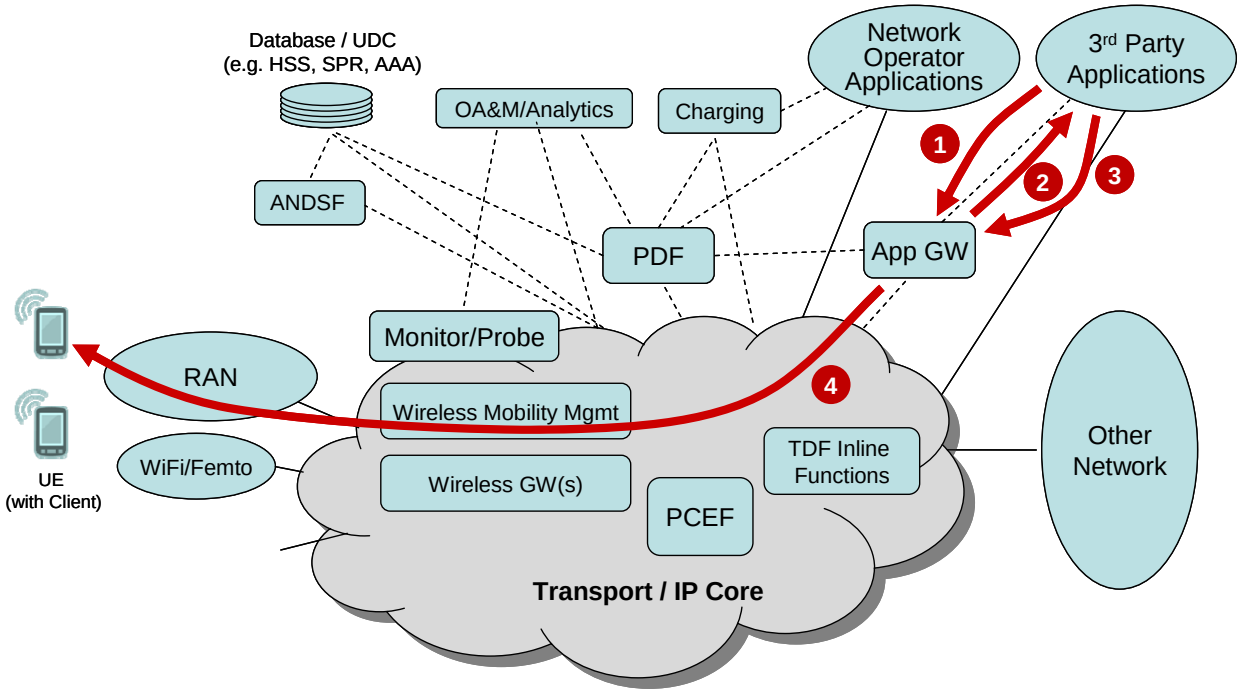

Figure 6.7.2-2: Network API Gateway for UE Status and Out-of-Firewall Push

1. When an application needs to check a user's connection status, the application sends a user connection status request to the AppGW using the network API
2. The AppGW retrieves the user connection status information from other network management nodes and sends the information to the application through the API.

3. When an application needs to push some notification to the user, the application delivers the push notification through an API to the AppGW, which contains some push delegation functions.
4. The network (AppGW) would identify when the user is in active mode and deliver the push notification to user.

A related option is to use the IETF Port Control Protocol (PCP) to allow an application client on the device to permanently open a pinhole in the network firewall. The Port Control Protocol allows an Ipv6 or Ipv4 host to control how incoming Ipv6 or Ipv4 packets are translated and forwarded by a Network Address Translator (NAT) or simple firewall, and also allows a host to optimize its outgoing NAT keepalive messages. PCP would increase the holding time on an IP address and may have security implications. Increased holding time may result in the need for larger pool of public IP addresses.

6.7.3 SDO/Gap Analysis

Some service capabilities referenced in section 6.7.1 are derived from 22.368 (Service requirements for Machine-Type Communications) and draft 22.801 (Study on non-MTC Mobile Data Applications impacts).

Because the massive machine-to-machine communication and smartphone usage are the major reasons for signaling-traffic congestion, there are several SDOs working on standards solutions to address the signaling traffic optimization problem.

3GPP is investigating the network optimization for M2M, which covers most of the service requirements in this section, in the following work items:

- 3GPP Network Improvements for Machine Type Communications (NIMTC) – Rel 10
- 3GPP System Improvements for Machine Type Communications (SIMTC) – Rel 11

3GPP is also investigating network optimization for better synchronization between network and applications for smartphone deployments in the following work items:

- 3GPP study on non-MTC Mobile Data Applications impacts (MODAI) – Rel 11
- 3GPP study on Interworking between Mobile Operators using the Evolved Packet System and Data Application Providers (MOSAP) – Rel 11.

3GPP has also created a study item on generic core network signaling overload prevention mechanism: 3GPP study on Core Network Overload (CNO) – Rel 11.

There are other activities related to signaling optimization, especially for M2M, in other SDOs, such as 3GPP2 Study for Machine-to-Machine (M2M) Communication for cdma2000 Networks (S.P0141).

There is still some standardization work needed to provide the optimized APIs between application and network/devices to reduce the signaling traffic. Messaging APIs should encourage the development of well-behaved and network-friendly applications. For this, work may need to be done in 3GPP/3GPP2 for the API between the application and network. OMA, Wholesale Applications Community (WAC) or other SDOs are responsible for the API between the application and the device.

7 SDO/GAP ANALYSIS SUMMARY

Areas for further standards development have been identified in the Section 6 use case analyses. Many of these identified gaps are common to multiple use cases. Figure 7-1 aggregates and summarizes all of these significant gap areas that are the focus of the standards-related recommendations of this report, and maps them to the applicable use cases enumerated below:

1. Congestion-aware Fairness
2. Subscriber-/Application-Aware Network Optimization
3. Network-Aware Scheduling of Content
4. User Rate Plans
5. Reasonable Network Protection and Management
6. Load- and Policy-Aware Multi-RAN Selection
7. Optimizing Use of Wireless Non-Bearer Resources

AREA	ITEM	Use Cases
Congestion Awareness	• Need better mechanisms to communicate congestion attributes to network control entities across the E2E network (e.g. radio access network load) • Need for continued analysis of and standardization efforts on CONEX and its alignment with Network Optimization	1, 2, 3
PCC - PDF	• Need for new interface on PDF/PCRF to provide network & subscriber state as inputs to policy control. • Extension to OCS-to-PCRF S_y interface needed, to use charging system for heavy user detection	1, 2, 3, 6, 7
PCC - TDF	• TDF enforcement actions should include a operator-defined application-specific user priority (optimization index) for inline TDF functions	2, 4
App/Network Interfaces	• Need to continue analysis of the standardization effort on API between application and network/device to reduce signaling traffic, including M2M. • Extensions to network APIs (OSA/Parlay-x/RESTful) to enable requests for reduced charging rate connections for content scheduling, and extension to the Rx interface and network APIs for a minimum re-try interval.	3, 7
Access Selection	• Recommend use of UDC for ANDSF policies to support better coordination with PCC subscriber policies. • Enhance ANDSF to allow the UE to utilize the access network availability information. • Need a mechanism to inform UE of coarse-grained access network availability at time of attachment.	6

Figure 7-1: Areas for Standards Development Summary

The following paragraphs provide a more detailed description of the above gap areas.

Congestion Awareness – Providing accurate and timely congestion indications as inputs and triggers for dynamic network optimization actions is a common need across many of the use cases analyzed in this study. Providing this information explicitly from key congestion points in the network (wireless RAN, shared wireline access, high-concentration aggregation points) will enable more accurate congestion indications and facilitate more valuable optimization use cases. Longer term, the IETF ConEx initiative holds the promise of a more end-to-end holistic

approach to managing congestion and aligning applications' priority, access to network resources, and network cost. As the ConEx standards develop, there is a need to ensure these network optimization use cases can take advantages of the richer end-to-end congestion indications, as well as policing and charging mechanisms.

PCC-PDF – Policy control is central to dynamic network optimization, and the Policy Decision Function is central to policy control. In order to make policy decisions that enable a wide range of optimization use cases, the PDF needs standard ways of obtaining inputs regarding both the state of the network, and the state of subscribers on the network. This requires a standard interface to the PDF, providing this information from other systems such as analytics and/or OA&M systems. In some cases, the subscriber information can be provided by charging systems. A standard OCS-to-PDF (Sy) interface exists, but would need to be extended in order for it to be used to indicate state changes – for example indicating subscribers' state changes in and out of a "heavy user" designation.

PCC-TDF – Traffic Detection Functions classify traffic in-line and can also apply optimization enforcements actions selectively to the various classified traffic flows. These enforcement actions can take many forms, and it may be impractical to standardize every specific enforcement action and the explicit control of those functions over the Gx interface from the PDF to the TDF. A practical approach would be to provide an application-specific user priority (or optimization index) over the Gx interface, which would allow operator-defined optimization enforcement actions (coupled to corresponding operator-defined policy rules) to be signaled between these two entities. This generic indexing method will require an extension to the existing Gx interface.

Application/Network Interfaces – Several optimization use case implementations utilize network APIs to interact between network control and applications and devices. One of the most important examples is the use of network APIs for minimizing the signaling impact of massive-scale M2M applications. Standard APIs need to be developed that will either replace or synchronize short duration but frequent signaling activity, such as heartbeat methods of determining device connection and health. Another specific example is the need for extending network APIs to allow reduced charging rate requests, as well as the need for extending network APIs and interface definitions on the Rx interface between PDF and Application Gateway to provide a minimum re-try interval (these as a way to support deferred scheduling of requested connection services and avoid APIs being overloaded by rapid re-requests when the network is too congested to satisfy the request).

Access Selection – It is recommended that ANDSF and PCC policies be stored in the UDC, to enable a consistent policy framework and coordination if needed.

It is recommended that some level (possibly coarse) of congestion/loading-level information be made available from wireless base stations to UEs in a way that can be utilized by UEs for access selection decisions at the time of network attachment, and correspondingly, ANDSF functions on the UE should be extended to take advantage of this congestion/loading-type information.

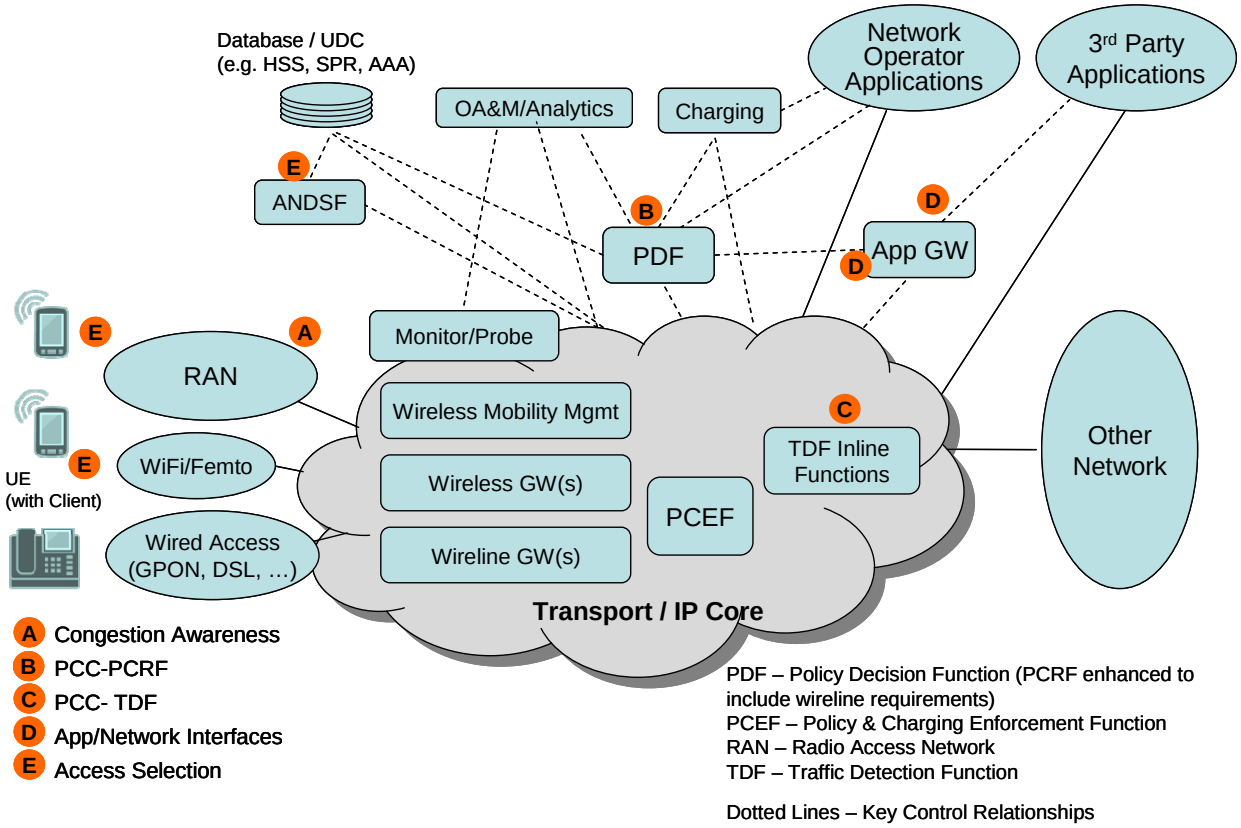

Figure 7-2: Areas for Standards Development Mapped to Architectural Framework

Finally, Figure 7-2 illustrates on the network optimization architectural framework introduced in Section 5, and used to analyze specific examples in Section 6, where the impacts are located for each of these standards gap areas.

8 LEGAL/REGULATORY SUMMARY

The previous sections of this document provide a technical analysis of Network Optimization with the goal of identifying key standards gaps impacting the architectures and use cases that may be used in the network. In the United States, there are important evolving regulatory requirements in the area of "net neutrality" to preserve Internet freedom and openness. This section discusses how the defined Network Optimization framework relates to current and proposed U.S. regulations pertaining to net neutrality.[9] This section discusses how the techniques proposed in this document relate to the regulatory framework. It is not intended in any way to provide a definitive assessment of the legality of any technique.

The FCC's *Open Internet Order*[10]addressing this matter was released on December 23, 2010. This order establishes four basic rules related to broadband internet access:

[9] This regulatory framework is subject to change via pending court decisions or by legislative action. This report should not be read to endorse or oppose the existing rules or any proposed changes thereto.

[10] *Preserving the Open Internet, Broadband Industry Practices*, GN Docket No. 09-191, WC Docket No. 07-52, FCC 10-201 (rel. Dec. 23, 2010)(*Open Internet Order*). Although the FCC released these rules in December, they have not yet

1. **Transparency.** Fixed and mobile broadband providers must disclose the network management practices, performance characteristics, and terms and conditions of their broadband services.
2. **No blocking.** Fixed broadband providers may not block lawful content, applications, services, or non-harmful devices; mobile broadband providers may not block lawful websites, or block applications that compete with their voice or video telephony services.
3. **No unreasonable discrimination.** Fixed broadband providers may not unreasonably discriminate in transmitting lawful network traffic.[11]
4. **Reasonable Network Management.** Fixed and mobile broadband providers may manage their networks flexibly and experiment with new business models, such as tiered pricing. Reasonable network management includes network management practices related to network security and integrity, network congestion, and additional activities identified through the Commission's case-by-case enforcement process. The Commission recognized that network management practices may differ between fixed and mobile networks, and it gave service providers the ability to seek out declaratory judgments that the proposed practices are reasonable.

In addition, the *Order* specifies that "[a] commercial arrangement between a broadband provider and a third party to directly or indirectly favor some traffic over other traffic in the broadband Internet access service connection to a subscriber of the broadband provider (i.e., "pay for priority") would raise significant cause for concern."[12]

It is also important to understand that the methods and use cases outlined in this document for network optimization can be applied to both internet services as well as managed or specialized services. The *Open Internet Order* does not apply the regulatory framework described above to such specialized/managed services unless such services, taken alone or together, are considered an *internet* service or a substitute for broadband internet access service.[13] Instead, the FCC will monitor the marketplace and take action as necessary to address any signs that specialized/managed services are harming competition or consumer choices (such as by retarding the growth of or restricting the capacity available for broadband internet service).

After careful consideration, we believe that the methods considered here can be made consistent with net neutrality rules. Furthermore, we believe the underlying structure of these optimization use cases, which are built upon the 3GPP PCC architecture, will also provide great utility for users and network service providers globally. It is very important that companies looking to implement network optimization methods seek guidance from their own regulatory counsels before implementing any mechanisms. The information provided below provides an overview of factors that may be considered but should not be construed as legal or regulatory advice.

---

become effective. The rules will become effective 60 days after a *Federal Register* notice announcing approval of the order's information collection requirements by the Office of Management and Budget.

[11] *Open Internet Order*, ¶1.

[12] *Id*. ¶76.

[13] *Id.* ¶114.

8.1 Congestion-Aware Fairness

The congestion-aware fairness use case applies traffic prioritization mechanisms to users associated with the congested part of the network based on a measure of fairness determined by whether the user is categorized as a heavy user or not. This determination is done programmatically and consistently by a network policy server.

This approach can be viewed as a variation of Comcast's "fair share" congestion management algorithm, which has previously passed muster with the FCC.[14]

The *Open Internet Order* specifically cites the congestion-aware fairness use case as an example of "Use-Agnostic Discrimination":

> Use-Agnostic Discrimination. Differential treatment of traffic that does not discriminate among specific uses of the network or classes of uses is likely reasonable. For example, during periods of congestion a broadband provider could provide more bandwidth to subscribers who have used the network less over some preceding period of time than to heavier users. Use-agnostic discrimination (sometimes referred to as application-agnostic discrimination) is consistent with internet openness because it does not interfere with end-users' choices about which content, applications, services, or devices to use. Nor does it distort competition among edge providers.[15]

8.2 Subscriber-/Application-Aware Network Optimization

This scenario is characterized by optimizations where a specific class or type of application is identified and further optimized (user "opt-in" mechanisms can be provided in any of these cases).

Take the example of a congested network with three flows – a video stream, a web browsing session, and an encrypted flow (e.g., using a virtual private network (VPN) or Pretty Good Privacy (PGP), or encrypted BitTorrent flows, etc.). In this example, a service provider could have two options for managing the congestion. One option is that the service provider could simply deliver what it could (i.e., limit all three flows to provide what it can). This works, but it may result in a lower QoE for some flows relative to other options. Another option would be to use different techniques on a per-application-type or per-content-type basis to provide improved quality of service where possible. For example:

- Use video compression on the video flow to reduce the bit rate but also allow the user to keep watching their video without having to experience video halts or restarts;
- Use jpeg compression to reduce the web session throughput; and
- As there is no option for intelligently optimizing the encrypted flow, this traffic would either continue to experience normal congestion losses according to its QoS priority class or it might be managed through rate policing.

---

[14] See IETF RFC 6057 Comcast's Protocol-Agnostic Congestion Management System, as well as Comcast's FCC disclosure document "FCC Network Management Response" (September 19, 2008).

[15] *Open Internet Order* ¶73.

The intent here is to employ reasonable network management during periods of congestion, so as to improve the aggregate QoE for the users (and their applications) via intelligent, application-aware techniques.

In examining this second option in the context of the *Open Internet Order*, there are three relevant areas of the rules to consider: blocking, non-discrimination, and reasonable network management practices. It would seem likely that this option would not violate the FCC's "no blocking" rule to the extent the video content, web browsing service, or encrypted application is not blocked or otherwise rendered "effectively unusable." Indeed, the intent here is to improve the user experience – where possible – during times of congestion. Recognizing that some congestion and packet loss are a normal part of any IP network, the FCC noted that its rule does not "impose a blanket prohibition on degradation of traffic more generally."[16]

The next hurdle concerns whether the content-specific optimization techniques could arguably violate the FCC's "no unreasonable discrimination" rule. While this rule initially appears to suggest that the differential treatment of traffic that discriminates among specific uses of the network or classes of uses could be discriminatory, the techniques are being applied with the intent of improving overall user QoE benefit during times of congestion. To the extent these reasonable network management practices are fully disclosed, end-users exercise "control" of the practices by consenting to the compression and throttling techniques (i.e. by explicitly opting in, for example as part of their terms and conditions of the service), and the practices in question conform "with best practices and technical standards adopted by open, broadly representative, and independent Internet engineering, governance initiatives, or standards-setting organizations," there arguably would be no violation of the FCC's "no unreasonable discrimination" rule.[17]

The FCC has not provided definitive guidance for broadband providers engaging in reasonable network management practices, noting that "providers should have flexibility to experiment, innovate, and reasonably manage their networks."[18] However, under its transparency rules, the FCC suggests – and we strongly concur – that it would be appropriate to disclose "whether and why the provider blocks or rate-controls specific protocols or protocol ports, modifies protocol fields in ways not prescribed by the protocol standard, or otherwise inhibits or favors certain applications or classes of applications."[19]

8.3 Network-Aware Scheduling of Content

In this scenario, the service provider is giving economic incentives to users to shift traffic to periods of low network congestion. This is not unlike the common practice of offering "night and weekend" rates for telephone service. So long as the incentives are disclosed to end-users, this mechanism would not seem to violate the *Open Internet Order*. The FCC has made it clear

[16] *Id*. ¶66, n.204.

[17] *Id*. ¶73.

[18] *Id*. ¶92.

[19] *Id*. ¶56.

that end-user choice and control are "touchstones in evaluating the reasonableness of discrimination."[20]

8.4 User Rate Plans

These use cases rely on user selection of specific rate plans, typically with different monthly usage volume caps and/or maximum throughput rates, to affect and manage network load to mitigate congestion. These use cases also seem to be consistent with the *Open Internet Order*. The FCC clearly states that its rules do not prevent providers from asking subscribers who use the network less to pay less and subscribers who use the network more to pay more, otherwise light end-users would be forced to subsidize heavier end-users.[21]

8.5 Reasonable Network Protection & Management

In this scenario, harmful and/or illegal traffic is targeted. The FCC has held that the service provider can apply network management techniques for: "ensuring network security and integrity, including by addressing traffic that is harmful to the network; addressing traffic that is unwanted by end-users (including by premise operators), such as by providing services or capabilities consistent with an end-user's choices regarding parental controls or security capabilities; and reducing or mitigating the effects of congestion on the network.."[22]

8.6 Load- and Policy-Aware Multi-RAN Selection

In this scenario, traffic is directed to one of several access options to optimize the traffic delivery. In considering this scenario relative to the FCC's rules, we note that:

- No traffic is blocked; and
- Relative to non-discrimination, this scenario can be implemented for all application classes or types. In some cases, the user experience is enhanced if different application classes use different offload policies, as long as all applications of a similar class or type are treated consistently. While the broadband provider would be treating specific uses of the network or classes of uses differently, end-users would control the practice by consenting to the multi-RAN selection capability, and the practices in question would conform to industry best practices and technical standards.[23] Therefore, there arguably would be no violation of the FCC's "no unreasonable discrimination" rule. Likewise, the FCC has indicated that it is less likely to find unreasonable discrimination when there is no harm to competition or end-users, which is another factor to consider in evaluating the practices in question.[24]

---

[20] *Id*. ¶71.

[21] *Id*. ¶72.

[22] *Id*. ¶82.

[23] *Id*. ¶¶70-71 and 74.

[24] *Id*. ¶75.

8.7 Optimizing Use of Wireless Non-Bearer Resources

This scenario targets traffic types and classes that may harm network operation. As noted above, the FCC has given broadband providers considerable latitude to engage in reasonable network management practices but has also suggested that it would be appropriate to disclose to end-users whether and why the provider blocks specific protocols, modifies fields in ways not prescribed by the standard, or otherwise inhibits or favors certain applications or classes of applications.

9 CONCLUSIONS & RECOMMENDATIONS

This study has explored a wide range of network optimization use cases, identifying the required service capabilities, various implementation options, regulatory considerations, and areas recommended for further standards development. While it is impractical to exhaustively address all use cases that fall within this wide ranging topic, the use cases that have been studied here are broadly representative of the network optimization that is of interest to the industry today—bringing real value to end-users, service providers, and network operators alike.

It is important to note that most of the use cases in this study rely upon policy management infrastructure to invoke and control optimization actions in the network. As a result, this work builds on the output of the ATIS Policy Management Focus Group (PM-FG), and extends it into application layer interactions.

Regulatory aspects are an important consideration in implementing any use case that alters the way the network serves the traffic demands placed upon it. This Report has looked at how use cases could be impacted by the FCC December 2010 order on net neutrality. After careful consideration, the team's conclusion is that the methods considered here can be made consistent with net neutrality rules, provided that proper opt-in mechanisms and transparent disclosure of network management methods are dutifully provided. It is important, however, that companies looking to implement network optimization methods seek guidance from their own regulatory counsels before implementing any mechanisms.

In the technical analysis of the seven use cases, there were several areas identified where further standards development is needed in order for these optimizations to be fully practical in multi-vendor networks. The specific development needs are summarized in Section 7. The NetOp Focus Group recommends actions with the applicable SDOs, as outlined below.

9.1 Shorter-term Standards Needs

9.1.1 ATIS Committees/FGs

Distribute the NetOp-FG report to all ATIS committees so they can take the NetOP-FG recommendations into consideration in their work program. In particular, the NetOp-FG has identified the following use cases that may be of interest to ATIS committees and Focus Groups:

- **M2M-FG:** The NetOp-FG "Optimizing Use of Wireless Non-Bearer Resources" use case deals with limiting the control plane impacts of M2M applications, which should be incorporated and extended in the work of the M2M Focus Group.

- **PTSC**: Several NetOp-FG use cases may be relevant to ongoing work in the PTSC, including Congestion Aware Fairness, Subscriber/Application Aware Network Optimization, and Reasonable Network Protection and Management.
- **PRQC**: The Congestion Aware Fairness use case may be relevant to ongoing work in the PRQC.
- **WTSC**: Several NetOp-FG use cases may be relevant to ongoing work in the WTSC, including Congestion Aware Fairness, Subscriber/Application Aware Network Optimization, Load and Policy Aware Multi RAN Selection, and Optimizing Use of Wireless Non-Bearer Resources.
- **CSF**: Several NetOp-FG use cases may be relevant to ongoing work in the CSF, including Subscriber/Application Aware Network Optimization, and Network Aware Scheduling of Content.

9.1.2 External SDOs

- **3GPP:** Many of the use cases in this study use network state, and especially mobile network congestion indications, as a primary factor for invoking network optimization actions. Recent initiatives within 3GPP to provide such inputs into the PCC architecture have not been accepted. It is one of the top recommendations of this focus group that ATIS member companies and the industry as a whole work toward a consensus on the value of exposing congestion information in the network and use of network congestion information in policy decisions (e.g., PCC and/or ANDSF) for optimization purposes.
- **OMA**: A number of use case implementations require APIs between the network and third party applications, and/or between the network and end-user devices. Further work in applicable SDOs such as the OMA is needed to detail the API functional needs of network optimization use cases. The NetOp FG analysis highlighted needs in the areas of deferred content scheduling and reduction of M2M signaling traffic.

9.2 Longer-term Standards Needs

- **IETF**: In the longer term, holistic congestion management approaches such as the IETF ConEx approach hold great promise for aligning application needs, user priorities, and network charging, policing, and overall congestion management in a comprehensive way. The NetOp-FG recommends that ATIS member companies actively encourage and advance the ConEx concept, especially in the areas of wireless RAN-specific aspects, as well as consistent and practical end-to-end congestion volume-based incentive/enforcement regimes.

It is recommended that member companies take a leadership role in drafting contributions and rallying support behind proposals to address these standards development needs and the other specific items outlined in Section 7.

APPENDIX A: ACRONYMS

Acronym	Definition
3GPP	3rd Generation Partnership Project
ANDSF	Access Network Discovery and Selection Function
AppGW	Application Gateway
APN	Access Point Name
BNG	Broadband Network Gateway
DIDA	Data Identification in Access Network Discovery and Selection Function (ANDSF)
DPI	Deep Packet Inspection
EMS	Element Management System
eNB	evolved Node B – an LTE basestation
GBR	Guaranteed Bit Rate
GW	Gateway
Gx	Gx is a 3GPP reference point that enables a PCRF to have dynamic control over the policy and charging behavior at a policy and charging enforcement point.
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IIF	IPTV Interoperability Forum
IMS	IP Multimedia Subsystem
IP	Internet Protocol
IPTV	Internet Protocol Television
ITU-T	International Telecommunications Union- Telecommunication
LTE	Long Term Evolution
MME	Mobility Management Entity
NAT	Network Address Translator
NMS	Network Management System
OCS	Online Charging System
OMA	Open Mobile Alliance
OPEX	Operational Expense
OPIIS	Operator Policies for IP Interface Selection (see 3GPP TR 23.853)
PCC	Policy and Charging Control

Acronym	Definition
PCEF	Policy Charging Enforcement Point
PCRF	Policy and Charging Rules Function
PDF	Policy Decision Function
PDN-GW	Packet Data Network – Gateway
PGP	Pretty Good Privacy
PM-FG	Policy Management Focus Group
PTSC	Packet Technologies and Systems Committee
QoS	Quality of Service
QCI	QoS Class Identifier
RAN	Radio Access Network
RNC	Radio Network Controller
Sd	Sd is a 3GPP reference point that enables a PCRF to have dynamic control over the application detection and control behavior at a TDF (Traffic Detection Function).
SDO	Standards Development Organization
SLA	Service Level Agreement
SPR	Subscriber Profile Repository: a logical entity that contains all subscriber/subscription related information needed for subscription-based policies.
Sy	Sy is a 3GPP reference point that enables transfer of information relating to subscriber spending from charging systems to PCRF
TDF	Traffic Detection Function
TOPS	Technical and Operations (Council)
UDC	User Data Convergence - 3GPP TS 22.101
UE	User Equipment
VPN	Virtual Private Network
WAC	Wholesale Applications Community

APPENDIX B: NETWORK OPTIMIZATION FOCUS GROUP MEMBERS

Chairs	
Marcus Weldon	Alcatel-Lucent
Paul Mankiewich	Juniper Networks
Members	
Ken Ko	ADTRAN
Kevin Sparks	Alcatel-Lucent
Ken Biholar	Alcatel-Lucent
Konstantin Livanos	Alcatel-Lucent
Farooq Bari	AT&T
Stephen Hayes	Ericsson
Jim McEachern	GENBAND
Amanda Xiang	Huawei
Tom Anderson	Juniper Networks
John Graybeal	Juniper Networks
Mike Fargano	Century Link
Kevin Hsu	Symmetricom
Naseem Khan	Verizon
ATIS Staff	
Tom Payne	
Paul Savitz	