New Wireless Broadband Applications and Devices: Understanding The Impact on Networks

MAY 2012
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TABLE OF FIGURES</strong></td>
<td>5</td>
</tr>
<tr>
<td><strong>EXECUTIVE SUMMARY</strong></td>
<td>7</td>
</tr>
<tr>
<td><strong>1 BACKGROUND</strong></td>
<td>9</td>
</tr>
<tr>
<td>1.1 MOBILE APPLICATION LANDSCAPE</td>
<td>9</td>
</tr>
<tr>
<td>1.2 MOBILE NETWORK ARCHITECTURE</td>
<td>10</td>
</tr>
<tr>
<td>1.3 INTRODUCTION OF UMTS AND LTE NETWORK ARCHITECTURE</td>
<td>11</td>
</tr>
<tr>
<td>1.3.1 LTE</td>
<td>13</td>
</tr>
<tr>
<td>1.3.2 Offloading Architectures</td>
<td>15</td>
</tr>
<tr>
<td>1.4 ACTIVITY STATES ON WIRELESS</td>
<td>18</td>
</tr>
<tr>
<td>1.4.1 The UMTS State Machine</td>
<td>18</td>
</tr>
<tr>
<td>1.4.2 The LTE State Machine</td>
<td>21</td>
</tr>
<tr>
<td>1.4.3 Interaction Between UMTS and LTE</td>
<td>22</td>
</tr>
<tr>
<td>1.5 QOS FOR APPLICATIONS, POLICY &amp; CHARGING</td>
<td>22</td>
</tr>
<tr>
<td>1.5.1 UMTS QoS</td>
<td>23</td>
</tr>
<tr>
<td>1.5.2 LTE QoS</td>
<td>24</td>
</tr>
<tr>
<td>1.5.3 QoS Management and Policy Framework</td>
<td>26</td>
</tr>
<tr>
<td>1.6 DEVICE PLATFORMS</td>
<td>27</td>
</tr>
<tr>
<td>1.6.1 iOS</td>
<td>28</td>
</tr>
<tr>
<td>1.6.2 Android</td>
<td>29</td>
</tr>
<tr>
<td>1.6.3 Windows Phone</td>
<td>29</td>
</tr>
<tr>
<td>1.6.4 Blackberry OS</td>
<td>30</td>
</tr>
<tr>
<td><strong>2 PROBLEM DEFINITION</strong></td>
<td>31</td>
</tr>
<tr>
<td>2.1 IMPACT OF APPLICATION BEHAVIOR</td>
<td>31</td>
</tr>
<tr>
<td>2.2 IMPACT OF DEVICE PLATFORM BEHAVIOR</td>
<td>33</td>
</tr>
<tr>
<td>2.2.1 Control Plane Impact</td>
<td>33</td>
</tr>
<tr>
<td>2.2.2 Data Plane Impact</td>
<td>37</td>
</tr>
<tr>
<td>2.3 TRENDS IN SECURITY</td>
<td>38</td>
</tr>
</tbody>
</table>
3.4.4 Network Technology Awareness ................................................................. 60
3.4.5 Lengthening Data Bursts ........................................................................... 60
3.4.6 Queueing up Data ...................................................................................... 61
3.4.7 Retry Scheduling ...................................................................................... 61
3.4.8 Text Compression and Media Transcoding ................................................ 62

3.5 ENHANCEMENTS WITH RESPECT TO SECURITY ........................................... 62
3.5.1 Malware Defenses ...................................................................................... 62
3.5.2 Packet Filtering .......................................................................................... 64

3.6 END USER BEHAVIOR ...................................................................................... 66
3.6.1 Increasing Awareness ................................................................................ 66
3.6.2 Incentives for Reducing Data Use .............................................................. 67
3.6.3 Equipping Users for Effective Usage Management .................................... 67

4 CONCLUSIONS .................................................................................................. 68

5 REFERENCES .................................................................................................... 69

ACKNOWLEDGEMENTS ......................................................................................... 70
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Separation of Service and Connectivity Layers</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>HSPA/LTE Network Elements &amp; Architecture</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>LTE/EPC Network Functional Architecture</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>Femtocell Architecture</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>Wi-Fi Offload architecture</td>
<td>17</td>
</tr>
<tr>
<td>6</td>
<td>The UMTS state machine (without URA_PCH or Cell PCH activated)</td>
<td>19</td>
</tr>
<tr>
<td>7</td>
<td>The UMTS state machine URA_PCH or Cell PCH activated</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>LTE State Machine has only two main states</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>Interaction between UMTS and LTE state machines (derived from 3GPP 36.331)</td>
<td>22</td>
</tr>
<tr>
<td>10</td>
<td>UMTS QoS architecture</td>
<td>23</td>
</tr>
<tr>
<td>11</td>
<td>Weekly application traffic for different subscriber clusters of a new Android smartphone model</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>(high-end with large screen) at one specific operator (Source: Ericsson Traffic And Market Data Report Nov 2011)</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Traffic Mix for data plans (Source: Ericsson Traffic And Market Data Report Nov 2011)</td>
<td>32</td>
</tr>
<tr>
<td>13</td>
<td>Global Sales of Devices by Type (in Millions) (Source: Strategy Analytics &amp; ABI Research)</td>
<td>33</td>
</tr>
<tr>
<td>14</td>
<td>Device-triggered Fast dormancy increases state changes and signaling when PCH state not is used</td>
<td>34</td>
</tr>
<tr>
<td>15</td>
<td>Signaling traffic generated by Apps</td>
<td>35</td>
</tr>
<tr>
<td>16</td>
<td>Push Service towards devices</td>
<td>36</td>
</tr>
<tr>
<td>17</td>
<td>Data Forecast from Cisco</td>
<td>37</td>
</tr>
<tr>
<td>18</td>
<td>Bad packet injection</td>
<td>39</td>
</tr>
<tr>
<td>19</td>
<td>Unconstrained growth in demand assuming 100% mainstream acceptance of leader behaviors. (This is an upper bound on the traffic forecast.)</td>
<td>42</td>
</tr>
<tr>
<td>20</td>
<td>Crossing the Chasm estimate of the demand curve without caps or Wi-Fi offload</td>
<td>42</td>
</tr>
<tr>
<td>21</td>
<td>US Daily Traffic forecast with and without expected Wi-Fi Offload included.</td>
<td>43</td>
</tr>
<tr>
<td>22</td>
<td>Average downlink data volume per subscriber versus device release date. From Michael Flanagan ©2012 by Arieso Inc., used with permission</td>
<td>45</td>
</tr>
<tr>
<td>23</td>
<td>Fitting curves to various historical data in Cisco’s VNI forecast</td>
<td>46</td>
</tr>
<tr>
<td>24</td>
<td>APIs are used to expose network capabilities to applications (e.g. OMA API)</td>
<td>49</td>
</tr>
</tbody>
</table>
Figure 25: for EUTRAN/EPC macro network (Fig. 5.6.3.2 in TS 23.829) ................................................... 53
Figure 26: Fast dormancy functionality ....................................................................................................... 54
Figure 27: Key LTE Radio Access Features ............................................................................................... 55
Figure 28: UE/device states on LTE ........................................................................................................... 56
Figure 29: High level overview of caching between device and application server ..................................... 60
Figure 30: Comparison of signaling, time taken for small bursts and large bursts ..................................... 61
Figure 31: Network Protection Architecture - Firewalls and IDP ................................................................. 63
Figure 32: Packet Filtering using Access Control Lists ............................................................................... 64
EXECUTIVE SUMMARY

Mass market mobile wireless applications and devices for the subscribers have exploded in recent years. The proliferation of new applications, devices and services has led to increased signaling and data traffic flowing through HSPA and LTE mobile broadband networks.

This white paper researches, analyzes and discusses some of the deeper aspects of addressing the increased load on HSPA and LTE mobile broadband networks due to the new types of traffic and increased Internet access on wireless. This includes identifying requirements on most parts of the ecosystem generating the traffic and making recommendations to help mitigate the effects of new devices and services.

- Section 1 of the paper sets the background and describes network architectures and principles. This section also introduces the reader to concepts and stakeholders identified later in the paper like QoS, device platforms and operating systems that influence network traffic.

- Section 2 provides the problem definition and introduces the reader to the effects that applications, devices and user behavior have on network traffic load. With the moving of primary Internet access from fixed to mobile devices, the changes to mobile device and infrastructure are also discussed.

- Section 3 builds up the previous sections to identify short term and longer term recommendations to address the growing mobile signaling and data traffic due to the explosion of devices and applications. These recommendations address stakeholders in the entire ecosystem, including application developers, mobile OS vendors, security on the mobile Internet, content providers and consumers, end users, etc.

The paper identifies some of the following challenges affecting network performance and end user experience:

- Impact of device platform behavior in both the control plane (always-on, constant push) and the user plane.

- New trends in device security arising from smartphone OS risks like malware.

- Impacts from Network trends like streaming, computing, storing, gaming, video.

Recommendations are made in some of the following areas:

- Enhancements to network architecture including optimizing service delivery environments, traffic management, and traffic offload solutions.

- Enhancements to device architecture to allow for optimization in radio resources, and network awareness.

- Guidelines for application development that consider the challenges of wireless access, caching, scheduling.

- Enhancements to address device security including Firewalls, intrusion detection, packet filtering etc.
• Education of end users in the areas of awareness of data/signaling usage, choice of appropriate access for heavy data usage, support for app data usage alerts etc.

Finally, some key conclusions of the paper include the following:

• Capacity for mobile broadband must continue to increase as new technologies and powerful devices that support bursty and chatty applications and drain network resources enter the market.

• Efficient use of radio resources through innovations for traffic offloading and the use of the best available access such as Wi-Fi, femtocells and picocells should be considered.

• Collaboration between network providers, application providers and device manufacturers is essential to address capacity needs and end user expectations. Improvements could be made to application design, device platforms and feature activations.
1 BACKGROUND

1.1 MOBILE APPLICATION LANDSCAPE

The rollout of high-speed wireless networks and devices was accompanied by the development of Wireless Access Protocol (WAP) browsers and device application platforms such as J2ME and Windows Mobile. Over time, networks have matured in capabilities and devices now feature larger, higher resolution screens and more powerful processors. This increased capability of the network and device hardware has motivated the development of robust software platforms incorporating HyperText Markup Language (HTML) browsers, comprehensive email clients and rich application platforms as exemplified by the iPhone, Android, BlackBerry, Windows Mobile, Symbian and WebOS smartphone platforms.

The advent of full HTML browsers on smartphones has made ubiquitous access to the web feasible and the development of WAP sites typically unnecessary. In addition, ubiquitous access to email has had a tremendous impact on productivity and has become an essential part of business. Typically most smartphone platforms today ship with integrated browsers and email clients that are optimized by the platform developer for both network utilization and user experience.

However, another big change in the use of mobile devices over the last few years has been the proliferation of applications or “apps” as they are most commonly known. These mobile applications offer convenient access to services through a rich user experience and are often free or have a minimal price tag. Further, the universal availability of application development environments from the application platform vendors and the ready availability of a distribution channel in the form of application stores have enabled the rapid development of the mobile application ecosystem driving significant use of mobile networks and a complementary impact on the use of mobile web browsing and mobile email usage as well.

Looking ahead, the next major step in the evolution of mobile applications is the transformation of voice and video calling into a mobile data application with the replacement of circuit switched voice and video calling by packet switched calling in next-generation networks. Further, as mobile browsers implement the full suite of HTML5 technologies, features characteristic of applications such as data persistence, native code execution and multi-tasking will become available to web sites/browser-based applications, effectively blurring the differences between browser-based applications and native applications. Another emerging trend is the use of always-on connectivity in conjunction with multi-tasking application environments to implement rich messaging services.

Driving the mobile application usage explosion has been the ready availability of a wide choice of innovative applications. These mobile applications are developed, delivered and monetized using mobile application platforms, which feature centralized stores or marketplaces for application distribution and billing integration to enable monetization. The mobile application platform consists of an operating system, application framework and key applications. A developer uses a Software Development Kit (SDK) corresponding to the mobile application platform, which gives them access to the necessary tools and Application Programming Interface (APIs) necessary to develop applications on the platform. The application framework enables reuse of components thereby enabling developers to build rich and innovative applications rapidly. In addition, the application framework enables developers to utilize features of the device hardware, access location information and run background services. Developers distribute their applications by submitting them to the application marketplaces where the marketplace operator may verify them before they are offered to end users. Further, the application marketplaces include review and rating mechanism for end users to provide feedback. These marketplace operator-
driven and end user-driven quality assurance mechanisms play a key role in the efficient operation of these marketplaces and the applications available through them. These mobile application marketplaces are operated by device manufacturers, service providers, mobile application platform vendors and other third parties.

The widespread uptake of smartphones, ready availability of application development platforms and distribution marketplaces are fostering innovation in applications on an unprecedented scale leading to an explosion in the number of applications, the usage of the applications and, naturally, the usage of mobile data networks.

1.2 MOBILE NETWORK ARCHITECTURE

Mobile networks are comprised of a number of network elements integrated to address the over-the-air communication over radio, the transport of data between the radio terminations and the Internet, the setup of communication sessions and enablement of services utilizing the data connectivity.

As shown in Figure 1, the elements of the network are:

- **Control or Signaling plane**: Enable the initiation and termination of communication sessions.
- **Data plane**: The elements of the network that enable the transport of the communication data over sessions, previously established using the Control or Signaling plane.
- **Connectivity layer**: Provides end-to-end data communication that may vary based on the type of session. For example, packet connectivity in HSPA does not involve the MSC, VLR; circuit switched voice in HSPA does not involve the SGSN, GGSN.
- **Services layer**: Translates the data communicated in the Connectivity layer into services. This is the abstracted functionality that utilizes the Connectivity layer to implement end-user friendly services.
The interaction between the Connectivity and Services layers is a focus area in this white paper. These would include APIs at the network level for communication with the Connectivity layer, and APIs at the device for the services/apps on the device to communicate with the Connectivity layer on the device.

Connectivity layer technologies in the Data plane include channel modulation, error correction algorithms and packet routing in conjunction with Authentication, Authorization and Accounting (AAA) and Quality of Service (QoS) elements in the network. Connectivity layer technologies in the Control plane include multiple access, duplexing and paging.

Service layer technologies in the Data plane include user level functionality such as voice, web browsing, videos, etc. with their own unique connectivity, content and interactivity requirements. Service layer technologies in the Control plane include Notification services for Application to Person (A2P) messaging.

The elements of the Service layer both in the Data and Control planes are typically made available to mobile application developers using standard software and web development abstraction paradigms such as networking protocols, sockets, markup languages, etc. Connectivity layer technologies in the Data and Control planes are typically reserved for management by the service providers.

The separation of the technologies employed across the mobile network into Connectivity and Service layers enables services to be developed independent of the underlying Connectivity layer technologies. Further, the layering enables mobile application developers to utilize the mobile network using paradigms familiar to them.

However, the limitations and capabilities of the Connectivity layer are not completely eliminated by the layering and an awareness of the underlying Connectivity layer in the Service layer enables improved performance of the total system and cross-layer optimization is the motivation for this paper.

Further, the availability of mature application or service development platforms both on client devices and in the cloud is triggering an exponential increase in the number of applications available on mobile networks. Although network operators continue to provide services within their domain, the subscribers are increasingly accessing the Internet for services as well. The Internet services encompass communication (email, VoIP, instant messaging), entertainment, social networking, navigation or other services. The growing trend is for many of these new services being developed to be delivered by application developers using platforms that are located entirely outside of the mobile network bringing with them unprecedented challenges for the Service providers. These challenges include a reduced ability to control the kinds of applications that make use of a network’s resources.

1.3 INTRODUCTION OF UMTS AND LTE NETWORK ARCHITECTURE

Universal Mobile Telecommunications System (UMTS) is a member of the family of third-generation systems defined by 3GPP, which combines innovations in radio access as an evolution from GSM. This is a new service architecture, which allows for mobile-fixed convergence at the service and application level. This network architecture is evolved into Long Term Evolution (LTE) networks, which bring efficiencies in the areas of latency, data rates and network architecture. See Figure 2.

UMTS:

- The system architecture of an UMTS network can be divided in two broad interacting domains: UMTS Terrestrial Radio Access Network (UTRAN) and the Core Network (CN).
The UTRAN consists of the Radio Base Station (NodeB) and the Radio Network Controller (RNC). Wideband CDMA technology was selected for the UTRAN air interface.

The main function of the CN is to provide switching, routing and transit for user traffic. The CN also contains the databases and network management functions. The UMTS CN is divided in circuit switched and packet switched domains. Some of the UMTS circuit switched elements are the Mobile services Switching Centre (MSC), Visitor location register (VLR) and Gateway MSC. UMTS Packet switched elements are Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN).

LTE:

The system architecture of an LTE network can be divided in two broad interacting domains: enhanced UMTS Terrestrial Radio Access Network (eUTRAN) and the Enhanced Packet Core (EPC).

The eUTRAN consists of the enhanced Radio Base Station (eNodeB). Orthogonal Frequency Division Multiplexing (OFDM) technology was selected for the eUTRAN air interface in the downlink and SC-FDMA (a CDMA technology) in the uplink.

The main function of the CN is to provide switching, routing and transit for user traffic. The CN also contains the databases and network management functions. The EPC consists only of a packet switched domain, with some common components from IP Multimedia Systems (IMS). EPC Packet switched elements are Mobility Management Entity (MME) and PDN gateway (PGW) and Serving Gateway (SGW).
For EPC, 3GPP (Third Generation Partnership Project) defines nodes common to EPC and IMS that will function in the different releases for future technologies. The Home subscriber server (HSS) acts for user control, and the Policy Control and Resource Functions (PCRF) specify access policies.

### 1.3.1 LTE

LTE describes standardization work by the 3GPP to define a new high-speed radio access method for mobile communications systems. This specification is standardized in 3GPP Release 8 (Rel-8). In this architecture, the main task is continuing with the smooth transition from 3G networks. LTE offers a smooth evolutionary path to higher speeds and lower latency. Coupled with more efficient use of operators’ finite spectrum assets, LTE enables an even richer, more compelling mobile service environment.

In order to achieve this, LTE networks (See Figure 3) offer enhanced:

- Spectrum efficiency
- Support for variable bandwidth
- Support for VoIP services based on continuous data connections
• Backward compatibility with pre-Rel 8 3GPP Releases
• Interfaces to interoperate with non 3GPP accesses e.g. CDMA and Wi-Fi accesses

One of the most significant features of LTE and EPC is its transition to a flat, all-IP based core network with a simplified architecture and open interfaces. Indeed, much of 3GPP's standardization work targets the conversion of existing core network architecture to an all-IP system. Within 3GPP, this initiative is called Evolved Packet Core (EPC) and was formerly known as System Architecture Evolution (SAE). EPC enables more flexible service provisioning plus simplified interworking with fixed and non-3GPP mobile networks.

Flatter network architecture in the EPC leads to improved data latency and better support of delay sensitive, interactive and real-time communications. EPC is based on IP protocols with support for TCP/IP, UDP, Diameter, SIP, etc. This allows operators to offer IP-based voice, video, rich media and messaging. This migration to all-packet architecture also enables improved interworking with existing fixed and wireless communication networks.

There are two nodes in the EPC architecture user plane; the LTE base station (eNodeB) and the EPC Gateway, as shown in Figure 3. This flat architecture reduces the number of involved nodes in the connections. The LTE base stations are connected to the core network over the S1 interface. Existing 3GPP (GSM and WCDMA/HSPA) and 3GPP2 (CDMA) systems are integrated to the evolved system through standardized interfaces, providing optimized mobility with LTE. For 3GPP systems, this means a signaling interface between the Serving GPRS Support Node (SGSN) and the evolved core network. For 3GPP2, a signaling interface between CDMA RAN and evolved core network addresses mobility.

Control signaling is handled by the Mobility Management Entity (MME) node, separate from the gateway, facilitating optimized network deployments and enabling fully flexible capacity scaling.
1.3.2 OFFLOADING ARCHITECTURES

Due to increases in smartphone usage and their high impact on 3GPP networks, analysis for new options to manage data traffic is a critical issue for mobile operators. Throughout the world and especially in the Americas region, there is a significant lack of spectrum.

Mobile data offloading is the use of complementary network technologies for delivering data originally targeted for cellular networks. Rules triggering the mobile offloading action can either be set by an end user (mobile subscriber) or an operator. The decision on choice of network usually resides in an end-user device, in a server or it is divided between the two. For the end-users, the purpose for doing mobile data offloading is based on data service cost control and availability of higher bandwidth. For the operators, the main purpose for the offloading is congestion of the cellular networks. The main complementary network technologies used for the mobile data offloading are Wi-Fi and femtocell.

1.3.2.1 FEMTOCELLS

A femtocell is a small cellular base station, typically designed for use in a home or small business. It connects to the service provider’s network via broadband (such as DSL or cable); current designs typically support two to four active mobile phones in a residential setting, and eight to 16 active mobile phones in enterprise settings. A femtocell allows service providers to extend service coverage indoors, especially where access would otherwise be limited or unavailable. Although much attention is currently focused on WCDMA, the concept is applicable to all standards, including GSM, CDMA2000, TD-SCDMA, WiMAX and LTE solutions.
For a mobile operator, the attractions of a femtocell are improvements to both coverage and capacity, especially indoors. Consumers benefit from improved coverage and potentially better voice quality and battery life. Depending on the carrier they may also be offered more attractive tariffs (e.g. discounted calls from home).

A femtocell-based deployment will work with existing handsets but requires installation of a new access point that uses licensed spectrum. In 3GPP terminology, a Home NodeB (HNB) is a WCDMA femtocell. A Home eNodeB (HeNB) is an LTE femtocell.

Typically the range of a microcell is less than two kilometers wide, a picocell is 200 meters or less, and a femtocell is on the order of 10 meters.

**Figure 4: Femtocell Architecture**

### 1.3.2.2 WI-FI OFFLOAD

Most data capable devices and smartphones now come with built in Wi-Fi capability. There are already thousands of installed Wi-Fi networks mainly in congested areas such as airports, hotels and city centers and the number is growing rapidly. One of the main concerns in a converged Wi-Fi/cellular network is to have a seamless and consistent user experience for subscribers. Wi-Fi offloading is an emerging approach to data offloading with multiple solutions entering to the market with using different capabilities in UE, external clients and SIM.

Standardization efforts have focused on specifying tight or loose coupling between the cellular and the Wi-Fi networks. 3GPP based Enhanced Generic Access Network (EGAN) architecture applies tight coupling as it specifies rerouting of cellular network signaling through Wi-Fi access networks. This makes Wi-Fi a de-facto 3GPP RAN.
3GPP has also specified an alternative loosely coupled solution for Wi-Fi. The approach is called Interworking Wireless LAN (IWLAN) architecture and it is a solution to transfer IP data between a mobile device and operator's core network through a Wi-Fi access. In this approach the two networks are in practice totally separated and network selection is done by a client application. In the IWLAN architecture, a mobile device opens a VPN/IPSec tunnel from the device to the dedicated IWLAN server in the operator's core network to provide the user either an access to the operator's walled-garden services or to a gateway to the public Internet. With loose coupling between the networks, the only integration and interworking point is authentication by an AAA connected to the HLR in the mobile network.

Very few terminals in the market handle IPSec connectivity in a native way today. Therefore, the UE needs an external client to support this functionality. The impact of installing this client and its behavior is a concern for new implementations.

The most straightforward way to offload data to the Wi-Fi networks is to have a direct connection to the public Internet. This no coupling alternative omits the need for interworking standardization. For the majority of the web traffic there is no added value to route the data through the operator core network. In this case the offloading can simply be carried out by switching the IP traffic to use the Wi-Fi connection in mobile client instead of the cellular data connection. In this approach the two networks are in practice totally separate and network selection is done by a client application.

Access Network Discovery and Selection Function (ANDSF) is the most complete 3GPP approach to date for controlling offloading between 3GPP and non-3GPP access networks (such as Wi-Fi). The purpose of the ANDSF is to assist user devices in discovering access networks in their vicinity and to provide rules (policies) to prioritize and manage connections to all networks.

![Figure 5: Wi-Fi Offload architecture](image-url)

With this final architecture, all policies and control defined by the mobile operator will be applied following business rules, and by the other side, all sessions will be handled by the same packet core providing continuity even if the user change his access from Wi-Fi to 3GPP networks and vice versa.

There are three main initiation schemes for mobile offload to Wi-Fi: WLAN scanning initiation, user initiation and remotely managed initiation.
1. In the WLAN scanning based initiation the user device performs periodically WLAN scanning. When a known or an open Wi-Fi network is found an offloading procedure is initiated.

2. In the user initiated mode a user is prompted to select which network technology is used. This happens usually once per network access session.

3. In the remotely managed approach, a network server initiates each offloading procedure by prompting the connection manager of a specific user device. Operator managed is a subclass of the remotely managed approach.
   - In the operator managed approach, an operator is monitoring its network load and user behavior. In the case of the forthcoming network congestion the operator initiates the offloading procedure.

### 1.4 ACTIVITY STATES ON WIRELESS

As the rapidly surging growth of smart phones in wireless networks, various types of applications including those “always-on” applications contribute as significant factors of consumption of UE power, bandwidth and network resources. Messages are sent between the phone’s active applications and their respective servers in the Internet, causing the phone to more frequently needing to connect/reconnect to the network, and subsequently moving between different states of the wireless state machine. The more frequent connections the phone needs with the network, the more it consumes battery, and the more frequent state changes between active and energy saving states, the more radio network management messages are needed, leading to increased signaling traffic in the network.

### 1.4.1 THE UMTS STATE MACHINE

In UMTS, the UE states supported are IDLE, CELL_FACH, URA_PCH, Cell_PCH and CELL_DCH. The last four states are for the UE in Connected Mode, that is, the UE that has established an RRC connection to WCDMA RAN. Different data rates are supported in CELL_DCH state. In the IDLE mode the UE does not have a logical connection with the radio network. The supported states are defined in the following sections. Further details on the states are specified in 3GPP TS 25.331.
Figure 6: The UMTS state machine (without URA_PCH or Cell PCH activated)

Figure 6 shows the transitions between the IDLE, CELL_FACH and CELL_DCH states.

In the IDLE mode, the UE is on and is able to receive system information and cell broadcast messages. It is not known to WCDMA RAN and cannot send or receive any user data. There is no RRC connection between the network and the smartphone. Power consumption for the idle state is lowest of all states.

In the CELL_FACH state, it is possible to transmit small amounts of low-speed data through the use of a shared channel, and in more recent standard releases (3GPP Rel-7, Rel-8), somewhat larger amounts of mid-speed data is supported as well. Connectivity is maintained while power consumption is about 20 to 30 times higher than in the idle state. The UE is able to transmit control signals and data packets on the common transport channel RACH or in common E-DCH (Rel-8) in the UL and on FACH or HS-DSCH (Rel-7) in the DL. No dedicated physical channel is allocated to the UE unless it is in the process of transmitting data on a common E-DCH. DRX patterns can further be applied to HS-SCCH channel to save battery energy for those UEs, which support this function.

The CELL_DCH state is characterized by the allocation of a dedicated connection over the air using DCH or E-DCH and by the allocation of HS-DSCH to the UE. The DCCH logical channels are used for control signaling and DTCH traffic channels are used for user data transmission. They are mapped onto the transport channels and further multiplexed onto one of Dedicated Physical Channels (DPCHs), DCH or E-DCH, or multiplexed onto shared physical channel, HS-DSCH.

1All logical channels are mapped onto RACH or E-DCH in UL and FACH or HS-DSCH in DL. These channels are suitable for carrying control information and with the Rel-7/Rel-8 enhancements some user data and the resources are shared by all users in the cell. No associated uplink HS-DPCCH is allocated if HS-DSCH is used, except while transmitting data on a common E-DCH in the uplink.
Figure 7: The UMTS state machine URA_PCH or Cell PCH activated

Figure 7 shows how state transitions are different when URA_PCH/CELL_PCH are used in the network and supported by the device.

In the URA_PCH state, the UE is known at the UTRAN Registration Area (URA) level in the Serving Radio Network Controller (SRNC) but it is not allocated with any dedicated radio resources and it is not able to transmit or receive any control signals or data packets. Essentially, the smartphone is in “sleep” mode, but since the RRC connection is maintained based on the most recent URA location, it enables the cellular network to still stay in a logical connection to the device, subsequently not breaking the application’s connection with their respective servers.

Similarly, in the Cell_PCH state, the smartphone is in “sleep” mode, but the RRC connection is maintained based on most recent cell location to enable the device to be directly addressed at its current cell when the network needs to wake up the device due to new data arriving. The device will update its location to the radio network whenever it moves from one cell area to another. Maintaining the UE context in the cellular network and not releasing the logical connection permits much faster network access than in idle mode, while still keeping power consumption as low as in Idle.

In Cell_DCH state a feature package called Continuous Packet Connectivity (CPC) improves the UE power consumption in that state (battery life time) and air interface capacity. The savings can be exploited either by enhancing the "always-on" experience by keeping inactive devices longer in the Cell_DCH state, or by further improving the device’s battery life time. UE DTX (discontinuous transmission from the UE) allows UEs to switch off continuous transmission of the dedicated physical control channel (DPCCH) when there is no information to transmit in the uplink. When this is the case, only a minimum of transmission is needed to maintain synchronization and control power. Two immediate benefits of switching off transmission are reduced battery consumption and reduced interference, which increases uplink capacity (in terms of number of users).
1.4.2 THE LTE STATE MACHINE

In LTE, 3GPP supports only two states for a UE\(^2\), namely, RRC IDLE, RRC CONNECTED. In the RRC CONNECTED state, there is a RRC context established – that is, the parameters necessary for communication between the terminal and the RAN are known to both entities. The cell to which the terminal belongs is known and the identity of the terminal is known, and these are used for signaling purposes between the terminal and the network. RRC CONNECTED is meant for data transfer to and from the UE. But discontinuous reception (DRX) can be configured to reduce power consumption, and long DRX operation is intended to achieve similar benefits as the PCH states in the UMTS system.

Although expressed differently in the specifications, RRC CONNECTED can be thought of as having two sub states, IN SYNC and OUT OF SYNC, depending whether the uplink is synchronized to the network or not.

---

\(^2\)3GPP TS 36.331 (LTE RRC States)
1.4.3 INTERACTION BETWEEN UMTS AND LTE

Figure 9: Interaction between UMTS and LTE state machines (derived from 3GPP 36.331)

Figure 9 addresses network and device transitions from UMTS to LTE networks. Devices in a connected state undergo in-session handover to the corresponding connected state in the other technology. In all other cases, reselection occurs between the non-connected states of the two technologies.

1.5 QOS FOR APPLICATIONS, POLICY & CHARGING

The explosion in data traffic in recent years can be attributed to the continuous growth and proliferation of a broad range of multimedia applications. While most users are not concerned with the details of how a particular service is implemented, users are interested in comparing the same service offered by different providers in terms of universal agreed upon parameters that focus on user-perceivable effects, rather than their causes within the network, and are independent of the specific network architecture or technology. Important parameters from a user perspective include end-to-end delay (including delays in the terminal, network and servers), delay variation due to the inherent variability in arrival times of individual packets in packet networks and throughput. Introduction of newer services such as VoIP, gaming, etc., will require strict guarantees on delay and packet loss in order for these services to be viable, thus highlighting the need for E2E (end-to-end) QoS (Quality of Service) mechanisms in the network.

QoS is a measure of the ability of the network to provide the necessary levels of service to the various applications and network flows. The spatial and temporal dynamics of radio environments coupled with the different error characteristics present unique challenges in meeting the necessary QoS demands needed to support the diverse end user applications.
There is a need to standardize simple and effective QoS mechanisms for multi-vendor mobile broadband deployments. Such QoS mechanisms could allow the access operator to enable service and subscriber differentiation and to control the performance experienced by the packet traffic of a certain service and subscriber group. To support QoS guarantees for various multimedia services, 3GPP clearly defines the end-to-end QoS architecture for UMTS and introduces several bearer and processing mechanisms to ensure UMTS can make full use of its technical advantages to offer differentiated services for users.

1.5.1 UMTS QOS

3GPP TS 23.107 specifies the UMTS QoS concept and architecture to best meet the requirements of end-to-end service flows in mobility networks. The UMTS QoS Architecture defines a layered QoS framework where there are multiple layers spanning different sections of the network between different nodes and is shown in Figure 10. An E2E service spans from one Terminal Equipment (TE) to another TE and is the top most layer of this architecture. The E2E service may have a certain QoS, which is provided for the user of a network service who decides whether they are satisfied with the provided QoS. A bearer service with clearly specified characteristics is set up from source to destination to meet the network QoS requirement. Bearer services provide QoS based on services provided by the layers below them that include TE/MT local bearer service, a UMTS bearer service and external bearer service. The bearer is the basic enabler for traffic separation because it provides differential treatment for traffic with differing QoS requirements. The concept of the bearer and the associated signaling procedures further enables the system to reserve system resources before packet flows that are mapped to that bearer are mapped into the system. The UMTS bearer service is comprised of the Radio Access Bearer service and the Core Network Bearer service, which, in turn, have sub-layers. The UMTS operator offers services provided by the UMTS bearer service that provides the UMTS QoS. Every bearer service must fulfill a set of QoS requirements.
UMTS QoS classes are defined by taking into account the different error characteristics of the air interface and accordingly, the QoS mechanisms defined must be robust. UMTS (3GPP 23.107) defines four different QoS classes namely:

- Conversational class
- Streaming class
- Interactive class
- Background class

These different QoS classes are distinguished from each other based on the delay sensitivity of the traffic supported by each.

Conversational and Streaming classes are used to serve real-time traffic flows, which are very sensitive to the delay. Examples of Conversational classes include real-time services like video telephony. Interactive class and the Background class are mainly meant for applications like WWW, email, FTP, News and Telnet. Since these classes are less delay sensitive compared to the conversational and streaming classes, both the classes provide better error rate by means of channel coding techniques and retransmissions. Packet retransmission is initiated whenever packet error, packet loss or packet order mismatch takes place because these classes expect high throughput and less error rates even though they are delay insensitive. The main difference between Interactive and Background class is that Interactive class is mainly used for interactive applications like interactive email and interactive Web browsing, while Background class is meant for background downloads or emails and background file downloading (3GPP 23.107). Since the scheduling algorithm gives more priority to the interactive class than the background class, the background applications use the transmission resources only when the interactive applications do not need them. Accordingly, performance targets for audio and video applications form the basis of network QoS classes. UMTS bearer service (QoS) attributes describe the service provided by the UMTS network to the user of the UMTS bearer. Key QoS service attributes (3GPP 23.107) include:

- Maximum bit rate (kbps)
- Guaranteed bit rate (kbps)
- Delivery order (y/n) etc.

The parameters related to throughput/bitrates are separated for uplink/downlink in order to support asymmetric bearers.

### 1.5.2 LTE QOS

3GPP LTE and SAE have enhanced the QoS mechanisms in UMTS. The evolved network, called Evolved Packet System (EPS), introduces the concept of default bearer that is setup when the user attaches to the network to effectively enhance user experience, reduce service setup latency and realize "always online" IP connections. The QoS parameters of the default bearer include the subscription data obtained from the Home Subscriber Server (HSS) whose values can be modified through PCRF interactions or local configurations. Other EPS bearers in connection with the same PDN are called dedicated bearers whose setup or modification can be triggered by the network side only. The main motivation for specifying the network-initiated QoS-control paradigm is that services are typically provided
by the access network operator. Therefore, it is natural that the access network and service owner assigns the QoS level per packet flow associated with a particular service. In addition, the bearer level QoS parameter values are always allocated by the packet core network. LTE supports end-to-end QoS where bearer characteristic are defined and controlled throughout the duration of a session between the mobile device (UE) and the P-GW. LTE QoS is characterized by:

- QCI (QoS Class Identifier)
- ARP (Allocation and Retention Priority)
- Maximum Bit Rate (MBR)
- Guaranteed Bit Rate (GBR)
- Aggregate Maximum Bit Rate (AMBR)

Bearer types are classified into two main classes: Guaranteed Bearers and Non-Guaranteed bearers, with guaranteed and non-guaranteed rates with QCI labels providing specific values of packet delay and loss that can be tolerated for any given bearer. The QCI value helps setup access point parameters used to control bearer level packet transfer (e.g. scheduling weights, admission thresholds, queue management thresholds, and link layer protocol configuration). ARP is used to decide the priority of admission or dropping off of dedicated bearers in case of limited network resources. Both GBR and non-GBR bearers are defined by QCI and ARP.

Besides QCI and APR, every GBR bearer is also associated with the parameters GBR and MBR. GBR bearers are used to carry voice, video and real-time gaming services through dedicated bearers. The GBR represents the bit rate that can be expected to be provided by a GBR bearer, while the MBR indicates the upper limit of GBR bearer.

Non-GBR bearers are used to carry best effort services and/or services whose bit rates cannot be guaranteed. To improve bandwidth utilization, EPS defines the AMBR. AMBR is an IP Connectivity Access Network (IP-CAN) session level QoS parameter of every PDN connection. Multiple EPS bearers for the same PDN connection share the same AMBR value. Each non-GBR bearer can potentially make use of the whole AMBR if other EPS bearers do not transfer any services. Theoretically, an AMBR restricts the total bit rate of all the bearers sharing this AMBR. AMBR can be classified into UE-AMBR and Access Point Name (APN)-AMBR. Table 1 lists standardized QCI characteristics defined in EPS.
Table 1: Standardized QCI characteristics: 3GPP TS 23.401 V8.1.0 (2008-03)

<table>
<thead>
<tr>
<th>QCI</th>
<th>Resource Type</th>
<th>Priority</th>
<th>Packet Delay Budget</th>
<th>Packet Error Loss Rate</th>
<th>Example Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GBR</td>
<td>2</td>
<td>100 ms</td>
<td>10^{-2}</td>
<td>Conversational voice</td>
</tr>
<tr>
<td>2</td>
<td>GBR</td>
<td>4</td>
<td>150 ms</td>
<td>10^{-3}</td>
<td>Conversational video</td>
</tr>
<tr>
<td>3</td>
<td>GBR</td>
<td>3</td>
<td>50 ms</td>
<td>10^{-3}</td>
<td>Real-time gaming</td>
</tr>
<tr>
<td>4</td>
<td>GBR</td>
<td>5</td>
<td>300 ms</td>
<td>10^{-6}</td>
<td>Non-conversational video</td>
</tr>
<tr>
<td>5</td>
<td>GBR</td>
<td>1</td>
<td>100 ms</td>
<td>10^{-6}</td>
<td>IMS signalling</td>
</tr>
<tr>
<td>6</td>
<td>Non-GBR</td>
<td>6</td>
<td>300 ms</td>
<td>10^{-3}</td>
<td>Video (streamed), TCP-based (chat, file sharing etc)</td>
</tr>
<tr>
<td>7</td>
<td>Non-GBR</td>
<td>7</td>
<td>100 ms</td>
<td>10^{-3}</td>
<td>Voice, Video Interactive gaming</td>
</tr>
<tr>
<td>8</td>
<td>Non-GBR</td>
<td>8</td>
<td>300 ms</td>
<td>10^{-6}</td>
<td>Video (streamed), TCP-based (chat, file sharing etc)</td>
</tr>
<tr>
<td>9</td>
<td>Non-GBR</td>
<td>9</td>
<td>300 ms</td>
<td>10^{-6}</td>
<td>Voice, Video Interactive gaming</td>
</tr>
</tbody>
</table>

To ensure interoperability between UMTS Terrestrial Radio Access Network (UTRAN) and Evolved UTRAN (E-UTRAN), an appropriate mapping between the EPS QoS Class Identifier (QCI) parameter and Universal Mobile Telecommunication System (UMTS) QoS parameter is designed.

1.5.3 QoS MANAGEMENT AND POLICY FRAMEWORK

The QoS management functions provide the functionality needed to establish, modify and maintain a UMTS Bearer Service with a specific QoS. The QoS management functions of all UMTS entities together shall ensure the provision of the negotiated service between the access points of the UMTS bearer service.

Policies are a set of rules that an operator can define and enforce in a telecom network, both within an operator domain and across different operator domains. The 3GPP standards have defined a Policy Charging and Control architecture that merges the service-based local policy and flow-based charging functionality into the model called PCC. Such a model allows for more efficient real-time control of the service flows in gateways.

In the PCC model, there are three key functional elements: a central policy repository, the policy decision point (PDP) and the policy enforcement point (PEP). Decisions made in the policy decision point (PCRF in 3GPP Rel-7), based on rules for network operations, are communicated to the policy enforcement points where these decisions are translated to QoS rules of individual service data flows to be enforced on the traffic plan.

The PCRF policy entity binds the service and transport layers. The PCRF collates subscriber and application data, authorizes QoS resources, and instructs the transport plane on how to proceed with the underlying data traffic. The PCRF is connected on its northbound Rx interface to the Application Function (AF). The AF resides on the service plane and represents applications that require dynamic policy and QoS control over the traffic plane behavior. The PCRF is connected on the traffic plane via the Gx interface to the Policy and Charging Enforcement Function (PCEF) that typically resides in a gateway node. The PCEF’s role includes traffic detection and resultant policy enforcement. A Subscriber Policy Register (SPR) node also provides subscriber specific data to the PCRF, to assist in evaluating policy decisions.
The service data flows can be thought of as a set of packet flows, typically IP flows and QoS control is applied per service data flow in the PCEF. The PCEF utilizes PCC (policy and charging control) rules to classify traffic by service data flow. Rules can be pre-defined or dynamically provisioned in the PCEF. Dynamic PCC rules are derived within the PCRF from information supplied by the AF (such as requested bandwidth), PCEF data (such as requested QoS at traffic level by user) and other Subscriber specific data if available. Provisioning of rules via the Gx interface to the PCEF can take place in two ways:

1. "Pushed" – i.e. unsolicited provisioning, where the PCRF may decide to provision PCC rules without obtaining a request from the PCEF; or

2. "Pulled“ – i.e. where Provisioning has been solicited by a request from the PCEF.

Each rule uses a series of data flow filters to allow the PCEF to detect the relevant traffic plane packets. The resultant activated PCC rule contains a QoS class identifier and the uplink plus downlink bit-rates authorized for the service data flow. As each PCC rule can only be bound to a single data bearer, this may require a series of rules to be installed to control QoS across multiple underlying traffic bearers.

The actual policy enforcement procedures for authorized QoS per PCC Rule is bearer dependent. Possible procedures include Packet scheduling, data packet (Diffserv) marking, and packet discarding. Event mechanisms can be set by the PCRF in the PCC rules to cause the PCEF to inform it of changes in the underlying traffic bearer.

1.6 DEVICE PLATFORMS

Device platforms have tremendous impact on not only how an application interacts with the device but also on how it communicates with the network. They are the foundational building blocks of a smartphone as they manage software resources and hardware interaction. Most smartphones in the market today have one of these popular Operating Systems (OS): Apple’s iOS, Google’s Android, Microsoft’s Windows Phone, RIM’s BlackBerry OS, HP’s WebOS, Nokia’s Symbian, and embedded Linux distributions such as Maemo and Meego. RIM’s Blackberry OS is a leading operating system amongst enterprise and business customers. Also, among these operating systems iOS and Android have gained popularity in the past three years with a collective market share of about 60 percent in 2011. Analysts have also predicted that Windows Phone will become the third most preferred OS by the year 2015 with a market share of approximately 20 percent. In this section we will focus on these four OSs and how they help smartphones efficiently interact with the network. The following table provides a snapshot of topics that are common to the four OSs, which will be covered in the subsections below.

---

3 http://www.gartner.com/it/page.jsp?id=1622614
4 http://www.gartner.com/it/page.jsp?id=1622614
### Table 2: Comparison of 4 major device OS

<table>
<thead>
<tr>
<th>Features</th>
<th>Apple iOS</th>
<th>Google Android</th>
<th>Microsoft Windows Phone</th>
<th>Blackberry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detection of network type</td>
<td>Yes via “Reachability”</td>
<td>Yes via ConnectivityManager</td>
<td>Yes via NetworkInterface Type</td>
<td>Yes via Network Manager</td>
</tr>
<tr>
<td>Asynchronous request/response</td>
<td>Supported by default</td>
<td>Supported by default</td>
<td>Supported by default</td>
<td>Supported by default</td>
</tr>
<tr>
<td>Data caching</td>
<td>Yes by default</td>
<td>Not supported by default</td>
<td>Not supported by default</td>
<td>Yes by default</td>
</tr>
<tr>
<td>Data Compression</td>
<td>Yes by default</td>
<td>Not enabled by default</td>
<td>Not supported by default</td>
<td>Yes by default</td>
</tr>
<tr>
<td>Supports network efficient data formats (JSON and XML)</td>
<td>Yes (iOS 5.0)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Push notification</td>
<td>Yes (4.0 and higher)</td>
<td>Yes (2.2 and higher)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### 1.6.1 IOS

iOS was developed by Apple Inc. and was first introduced in 2007 when the first iPhone launched. Apps for the iOS can be developed using the Cocoa Touch framework, which is implemented using objective-C. Web applications that run on the Safari browser are built using Web2.0 technology. iOS offers many features that help its devices and their applications to efficiently utilize network resources and provide a good end-user experience. An application has to be aware of the current network status to determine if data packets can be routed through the cellular data network. If no network connection is available, or if the signal strength is too low to sustain the data transfer, the application can switch to offline mode to preserve battery life in the connected state. iOS provides a sample application called “reachability,” which can be used to determine the network status. The advanced Safari browser included in iOS supports the latest HTML5 offline data storage features. The offline storage means a web app can store session data locally in a cache on the device, using either a simple key/value data API, or a more advanced SQL interface. The data is persistent among Safari launches, meaning apps start up faster, are less dependent upon the network, and perform better. iOS supports data compression formats such as gzip and deflate and automatically adds “Accept-Encoding” header in the requests and decompresses the response. This helps in saving the amount of data that needs to be transferred over the network for network enabled applications. Apple also introduced support for push notification in iOS 4.0. Thus an application is not required to keep polling the server for updates or run in the background. If an update for an event occurs,

---


the server would send a push notification to the application. Starting with iOS 5.0 Apple supports native parsing for lightweight data formats such as JSON without having to use third party libraries. The OS handles HTTP threads asynchronously so that a network error in sending a request or receiving a response does not block the main thread preventing the application from becoming non-responsive. There are third party libraries like Three20 and ASIHTTPRequest that support handling of multiple simultaneous requests and responses independent of each other\(^7\). The advantages of asynchronous request and response are detailed in section 3.3. Apple also joined efforts with Nokia to introduce “Network Controlled Fast Dormancy” in iOS 4.2. This technology helps cut down signaling traffic and increase battery life for the device. Network controlled fast dormancy and its advantages are detailed in subsequent sections.

### 1.6.2 ANDROID

Android is a Linux-based open source OS developed by Open Handset Alliance (OHA) led by Google. Applications for Android are written in Java\(^8\). The OS supports asynchronous handling of multiple HTTP threads and completion of tasks in the background mode. The main UI activity is kept separate from other threads so that main UI is not blocked due to pending operation on other threads. Android provides ConnectivityManager to determine the availability and type of network connection. Thus an application can check the connectivity status before starting the update of a scheduled background service or data cache. The OS also provides techniques to delay the download of bandwidth intensive applications or data until the device is connected to Wi-Fi\(^9\). This Wi-Fi offloading facility takes the burden away from the cellular data network. HTTP caching is supported in Android and its persistent cache size is 6MB. The WebKit browser in Android supports a cache memory of 8MB\(^10\). Like iOS, Android also supports data formats such as JSON and XML for text encoding. These formats are supported by most web servers and are easier and faster to parse and present results to the end user. Starting version 2.2, Android supported Cloud to Device Messaging (C2DM), which enables servers to send push notification to the device whenever fresher data becomes available. This increases network efficiency as an application does not have to wake up periodically and poll the server for data, where, sometimes, new data might not be available. This also helps preserve battery life of the device. Android supports data compression formats such as gzip and deflate but compression is not enabled by default.\(^11\) In conclusion, Android, due to its openness nature, has left it to the developers to come up with solutions to make their apps network efficient and does not offer concrete guidelines on how to build network friendly applications.

### 1.6.3 WINDOWS PHONE

Windows phone, created by Microsoft, is the evolution of the Windows Mobile platform that began with Windows CE in 1996. Microsoft released the SDK to developers in September 2010. Development for Windows Phone application is done using the following Visual Studio 2010 (or higher), expression blend for design purposes, C#, .Net, Silverlight for business applications, and XNA for game development. As with iOS and Android, this OS also supports asynchronous handling of network requests and response out of the box. There are two classes supported in .Net framework that handle asynchronous

---

\(^7\) GSMA - Smartphone Challenge: Guidelines for development of network friendly applications  
\(^11\) GSMA - Smartphone Challenge: Guidelines for development of network friendly applications
request/response for accessing data from the Internet: WebRequest class and HttpWebRequest class. The NetworkInterfaceType property in the NetworkInterface class provides the network type servicing Internet requests. However this information is not available instantaneously and therefore Microsoft recommends checking network connection type on a background thread. The OS does not support caching by default. Caching of network data is only partially supported by the Silverlight engine. An example of how an offline data cache can be implemented is available. Memory, storage and communication bandwidth conserving data formats such as JSON and XML are supported by the OS. However serialization using the above data formats comes at the cost of increased processing resources. As with Android this OS does not support compression by default. The industry uses multiple software libraries to perform decompression for content that is gzip-compressed by a web server. The platform provides a push client service to enable push notifications by default. A comprehensive explanation of the service is explained here. It is rumored that Windows Phone 8, code named “Apollo” will have features such as data smart, which will help users save cellular data when possible and avoid “bill shock,” and Wi-Fi offloading, which detects Wi-Fi networks and connects to them automatically without requiring any user interaction.

1.6.4 BLACKBERRY OS

BlackBerry 10 is a multitasking, microkernel operating system built upon the QNX Neutrino RTOS, which has a long history in the automotive, industrial, medical and networking markets. BlackBerry 10 runs device drivers, file systems, networking stacks and applications outside of the OS kernel as memory-protected components.

This modular architecture enables the design of self-healing systems that can recover gracefully from software faults in applications or system-level services; it also allows systems to support new or upgraded services on the fly. The BlackBerry 10 networking stack executes outside the kernel, like the other system-level processes, and presents developers with a single unified interface, regardless of the configuration and number of networks involved. Its active networking executable implements a zero-copy architecture, which provides the best possible performance for network interaction. Based on the open NetBSD, the networking stack supports all the more common protocols and networking functionality: UDP, TCP, IP (v4 and v6), broadcast, multicast, forwarding, routing, sockets, multilink PPP, PPPoE, DHCP, DNS, NFS, NTP, IPsec, and so on.

The web browser for BlackBerry 10 is based on WebKit, an open source project and the most popular web browser layout engine, used in Apple Safari, Google Chrome, Amazon Kindle eBook reader, and many other browsers. The browser for BlackBerry 10 enables true desktop-style browsing on smartphones and tablets, with Adobe Flash support and industry-leading HTML5 compatibility. Its hardware-accelerated features include CSS3 animations and transitions, 2D and 3D canvas, as well as embedded 1080 pixel HTML5 and Flash video, and smooth 60 fps touch scrolling. The browser incorporates a regular, up-to-date WebKit base for the latest features and defect fixes.

---

13 GSMA - Smartphone Challenge: Guidelines for development of network friendly applications
2 PROBLEM DEFINITION

2.1 IMPACT OF APPLICATION BEHAVIOR

The reality of user behavior in consuming more content than what is produced, the popularity of multimedia services and the lack of processing power on UEs results in far greater traffic flowing downstream into UEs than flowing upstream from UEs into the cloud. Naturally, communication channels are typically asymmetric reserving greater bandwidth for the downlink vs. the uplink. In spite of the asymmetric bandwidth allocation, mobile networks are seeing significant traffic pressure on the downlink due to the sheer number of applications using the networks and the multimedia heavy nature of the traffic generated by many of the applications.

Figure 11: Weekly application traffic for different subscriber clusters of a new Android smartphone model (high-end with large screen) at one specific operator
(Source: Ericsson Traffic and Market Data Report Nov 2011)

As shown in Figure 11, the dominant contributors to mobile traffic are online media and web browsing, which tend to be extremely asymmetric with data primarily flowing downstream from the cloud into the UEs. Another emerging class of traffic on the downlink is OS and application updates, which are delivered over-the-air to the UEs.

---

In addition, a number of symmetric applications such as file sharing or Peer-to-Peer (P2P), mobile Voice over IP (VoIP) and gaming contribute to traffic on the downlink. However, P2P traffic has been on the decline over the last several years as first audio content and then video content became readily available through legitimate storefronts. In addition, relative to other traffic types, Mobile VoIP and Mobile Gaming are low bit rate services, which do not impact the overall traffic as much in spite of a very large number of concurrent sessions in use.

As UEs become more capable, services such as video conferencing, User Generated Content (UGC) uploads, P2P applications, video surveillance and augmented reality are gaining popularity. Further, as voice services transition to VoIP in LTE networks, the potential increased traffic from voice will also contribute to the uplink traffic mix. The aggregate of these services is making network capacity requirements more symmetric and resulting in significant pressure on the uplink capacity. Foreseeing this trend, HSPA and LTE have increased the relative throughput of the uplink vs. the downlink compared to earlier technologies\(^1\) and further LTE also allows for more flexible spectrum allocation to account for evolving network traffic usage patterns.

Nevertheless, until LTE is widely deployed, accelerated growth in uplink traffic will require immediate and unique solutions on deployed HSPA and HSPA+ networks.

\(^{16}\)HSPA Evolution — Boosting performance of mobile broadband access
(http://www.ericsson.com/ericsson/corpinfo/publications/review/2008_01/05.shtml)
As shown in Figure 13, Machine-to-Machine (M2M) devices are expected to grow into a significant share of the devices on the network. These M2M applications have diverse network requirements varying from heavy signaling-low throughput in the case of geo-tracking to low signaling-high downlink throughput in the case of Near-Video On Demand (VOD) to low signaling-high uplink throughput in the case of webcams. Further, M2M applications may also be widely distributed in large numbers due to the low cost-low maintenance nature of the UEs resulting in rapid growth in M2M traffic. In addition, the wide distribution of these devices will require remote manageability using OTA software updates further adding to the traffic demands on the network.

### 2.2 IMPACT OF DEVICE PLATFORM BEHAVIOR

As the 3G service gets widely deployed and next-generation service starts to kick off, various types of mobile devices are introduced into the market. Today’s mobile devices are multi-functional devices capable of hosting a broad range of applications for both business and consumer use. Providing this ever-increasing capability is the mobile operating system, which is also evolving at a rapid pace, including the provision of frequent updates or upgrades to the end users for downloading to their mobile device. As discussed further in this section, this can cause an impact both to the mobile device itself, such as reduced battery life, and network performance degradation, such as signaling storm.

#### 2.2.1 CONTROL PLANE IMPACT

As the rapid growth of various mobile devices being supported by UMTS/LTE network continues, network congestion becomes one of the biggest issues, highlighted by the popularity of smartphones and M2M usage. However, the capacity crunch is not only due to increasing user data traffic, but also in increasing signaling traffic. Due to the fact that the applications that run on smartphones and M2M have different behavioral characteristics compared to that of traditional voice and simple data, which run on the wireless network, they pose a large signaling traffic challenge to the networks which support. For example, smartphones and M2M have an increasing number of applications that send only a small amount of data, but the transmission frequency of the packets is relatively high. Users of smartphones make constant queries to the network as they move among cell sites to push email, access social networking tools and conduct
other repetitive actions. These always-on applications also rely on keep-alive messages. As a result, while data traffic is growing, by many accounts, signaling traffic is outpacing actual mobile data traffic by 30 to 50 percent, if not higher. As another example, a web-based IM user may send a message but then wait a couple of seconds between messages. To preserve battery life, the smartphone moves into idle mode. When the user pushes another message seconds later, the device has to set up a signaling path again. The base station controller (i.e., RNC) spends a lot of its resources trying to process such signaling, which prevents it from doing other things like allocating additional resources for data.

There are several common smartphone behaviors which could potentially bring signaling storm to the network and need to be considered and handled within the ecosystem. Some key behaviors are:

- Fast dormancy
- Heartbeat for always-on application
- Constant push service
- Network (re-)attachment

2.2.1.1 FAST DORMANCY NETWORK IMPACT

It is essential to keep the UE power consumption low while having to do frequent transmissions. The original idea in 3GPP was to provide the Radio Resource Control (RRC) states Cell_PCH and URA_PCH to allow for very low UE power consumption. Here, when a data transmission is over, the UE is moved to PCH state quickly and there would be no need to move UE to idle state for low power consumption. The motivation to avoid idle state is because the next packet would then cause packet connection (PS RAB) setup leading to increased latencies and increased signaling traffic in the network.

![Figure 14: Device-triggered Fast dormancy increases state changes and signaling when PCH state not is used](Source: Ericsson presentation on impact of Smartphones on 3G network, 2010)
Nevertheless, it is common practice that many networks are configured with relatively long inactivity timers for Cell_DCH and for Cell_FACH states, resulting in infrequent transitions to the PCH state. As a result, transmissions of even a small packet may lead to high battery drain. In order to keep UE power consumption low, UEs have implemented proprietary (that is, a functionality not defined by 3GPP standards) Fast Dormancy. When using such a Fast Dormancy, the mobile application informs the radio layers when the data transmission is over, and the UE can then send Signaling Connection Release Indication (SCRI) to the RNC simulating a failure in the signaling connection. Consequently, the UE releases RRC connection and moves to idle state. That approach keeps the UE power consumption low, but it causes frequent setup of packet connections unnecessarily increasing network signaling load. There are also differences between different UE vendors on how they implement Fast Dormancy functionality. In addition to the high signaling load, the network counters indicate a large number of signaling connection failures as this battery saving method cannot be distinguished from a genuine signaling connection failure in the network.

### 2.2.1.2 HEARTBEAT FOR ALWAYS-ON APPLICATION

One of the key driving forces of the smartphone is the huge number of online applications available to its users and its ability to provide a mobility platform for user to access those applications and information anytime and anywhere. For real-time Internet applications, the logical always-on connection between the server and client is required, resulting in frequent or periodic small heartbeat packets to be sent as a keep-alive message to maintain the connection. Since there is no coordination between the application and network on when these small heartbeat packets will be sent, and considering that mobile devices are frequently switched to idle mode to save battery, many times these heartbeat packets end up being sent during the mobile device’s idle state, which, in turn, triggers the device to continually switch between active and idle model over short period time and create massive signaling traffic. Although heartbeat is mainly associated with a given application and can be considered as having impact to the application itself, the smartphone platform becomes tightly coupled with many always-on applications which further contribute to the heartbeat signaling traffic. The following figure demonstrates signaling traffic created by different platforms. In addition, the smartphone platform also has its own heartbeat mechanisms to keep the device connected to the server all the time.

![Figure 15: Signaling traffic generated by Apps](Source: Huawei Smart Lab)
2.2.1.3 CONSTANT PUSH SERVICE

Currently the major smartphone platforms support push notification mechanisms to allow third party applications to use device platform vendors’ push server and push mechanism to convey some short notification information to the client in the device. This is done in order to let the user receive the latest update or information from the application server. Some applications running on particular device platform can create more push related signaling traffic. Since these push servers reside outside the operator’s network and are not coordinated with the device’s actual network connection status, such as idle or active mode, many push messages may therefore be sent while the device is on idle mode, which will trigger unnecessary signaling traffic, such as paging messages, network connection and release messages.

![Push Service](image)

Figure 16: Push Service towards devices

2.2.1.4 NETWORK ATTACHMENT

Since the notion of “always online” is critical to the end user of smartphones and the smartphone is becoming the device of choice using operators’ next-generation mobile broadband networks, many mobile devices are designed to aggressively keep connecting to the broadband network as early as possible to ensure a good user experience. However, due to different device platform implementation, different level of signal traffic impact can be expected. For example, while some operating systems may establish data session context as soon as their device is powered on, others may only do so when a certain application is launched. Each will obviously have a different impact on the network, with the actual impact being dependent on the penetration rate of each operating system in a given market.

Similarly, in response to scenarios involving a network failure, where an IP connection is lost or a device fails to attach or reattach to the network, some device platforms will conduct immediate and consistent network attach retries until the device is able to attach to the network. Large concentrations of mobile devices using platforms behaving this way could potentially trigger a signaling storm.
In addition, currently almost all smartphones support multiple RAT to make sure the user has seamless service experience and has the option to select different broadband connections. Considering the coverage limitation of the initial LTE deployment, some potential frequent inter-RAT handover, network exit and re-entry are expected, especially in the overlay area of multiple RATs, therefore the signaling traffic associated with those activities can dramatically increase.

2.2.2 DATA PLANE IMPACT

According to Cisco’s latest mobile traffic forecast report, Smartphones represent only 12 percent of total global handsets in use today, but they represent over 82 percent of total global handset traffic.

Modern mobile device platforms are in a continual state of evolution even after deployment in the field. This requires updating of the firmware and software on these devices to be delivered over the air both to provide new functionality and also to address bugs and security issues in the deployed firmware and software. In addition, open device platforms also lead to the proliferation of malware such as Trojan viruses, viruses and behavioral probes that require UE-based security scanning software that need to be continually updated with malware signatures. The frequency of these updates and the need to update every single device when an update is required, especially to ensure security of the network, is a growing contributor to downlink traffic.

Because the smartphone is not only the communication platform but also more an application platform, the device platform is continually evolving in order to provide more functionalities to facilitate the applications running on them, and these new functions are creating more uplink and downlink traffic.
As operators are facing pressure to keep their HSPA/LTE networks to meet the bandwidth demand by the smartphone and M2M users, Wi-Fi offloading is becoming a supplement to HSPA/LTE network, and all of the smartphones have Wi-Fi built-in. But there are differences between different device platforms on how to handle the user traffic between Wi-Fi and HSPA/LTE network. While some platforms may support simultaneous Wi-Fi and HSPA/LTE data, selecting certain data for transmission over one or the other, other platforms may disable HSPA/LTE connection as soon as Wi-Fi connection is established.

### 2.3 TRENDS IN SECURITY

#### 2.3.1 SECURITY RISKS ASSOCIATED WITH SMARTPHONE

While open-device platforms allow for a broad range of innovative smartphone applications for the end user to choose from and utilize, they also can be exploited to create potential security risks to the network. However, while over 2,500 different types of malware had been identified in 2010, only 5 percent of smartphones had some sort of security software in place as of October 2011. With the rapid increase of smartphone usage, increase of mobile network applications, and lack of security protection being utilized, smartphones are considered to be a source of different types of security risks and threats on mobile networks. In the next sections we will discuss a couple of the more currently prevalent potential security threats.

#### 2.3.2 MALWARE PROPAGATION

The major threat of malicious software, or malware, on mobile networks is the resulting increase in traffic, consuming and wasting precious resources and bandwidth of both the air interface and core network of mobile networks. Historically, the impact of malware has been primarily to the IP terminal itself, such as the personal computer. With mobile networks, however, the threat is potentially more widespread, as the downloading of the advertisements to an IP terminal such as a smartphone will waste air interface resources to do so. One well-known example is adware, which is software that automatically displays advertisements on the IP terminal. Spyware is another type of malware, which monitors the activities of the user and reports those activities to a malicious server, also wasting resources and air interface bandwidth.

#### 2.3.3 BAD PACKET INJECTION

Modern mobile devices such as smartphones are also capable of injecting bad packets into the mobile network, which might create denial-of-service attacks on supporting network elements. In these scenarios, rouge mobile device(s) would have the ability to insert fake packets. The presence of these packets could then lead to unnecessary additional load on network elements and/or misrouting/loss of intended packets; with the severity of the impact being related to what extent the bad packet is propagated throughout the network.

---

2.4 TRENDS IN NETWORK

Much depends upon forecasts of future demand for wireless services, particularly in light of the extraordinary growth rates seen in the recent past few years. Recent analyses have characterized the last several years of mobile data volumes as doubling each year implying 700 percent growth in three years, reflecting that the demand, while surging, is moderating somewhat\(^2\). For operators to properly set pricing and marketing parameters as well as procure appropriately engineered Radio Access Network equipment, one must thoughtfully forecast future demand.

---

Forecasting can be done with many varied approaches including:

1. A detailed assessment of different types of subscribers and their specific application use patterns and extrapolate each type of user and application separately

2. Sociological and psychological assessment of evolving usage patterns across various demographic groups

3. Anticipating new services such as Machine-To-Machine applications with very different usage patterns and willingness-to-pay

4. Consideration of historical analogs such as fixed Internet access which the wireless industry has been trailing with a few years delay

5. Simple extrapolations from model fit to historical data

It is important to get these forecasts right and to keep them refreshed so as not to over-invest as in the “dot com” boom. Likewise, under-investing a service can leave an operator open to competitive threats and missed opportunities. These are billion dollar decisions that require all these various approaches to forecasting the critical market drivers.

2.4.1 ASSESSMENT OF SUBSCRIBERS AND APPLICATIONS

As a bound, it is useful to consider, as in approach 1 and 2 above, what the most demanding users for the various application types would be and extrapolate what would happen if the entire population adopted their use patterns. The pragmatists follow the visionaries in “Crossing the Chasm.”

The “five big apps” and their most ardent user demographics are shown in the following table. We also suggest possible enablers for mainstream adoption, far beyond the existing leaders in the use of these applications. We also suggest what level of traffic these leaders may demand in 2016. The traffic demands are for total traffic including 3G, 4G and Wi-Fi smart loading, and they reflect the unconstrained traffic demands of these leading users.

---


### Table 3: Major Applications and User Demographics

<table>
<thead>
<tr>
<th>Application</th>
<th>Definition</th>
<th>Examples</th>
<th>Demographic</th>
<th>Mainstreaming enablers</th>
<th>Leader’s traffic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Streaming</strong></td>
<td>Audio or video streaming</td>
<td>YouTube, Hulu, Netflix, Livestream, monitoring &amp; health surveillance.</td>
<td>Mid-lifer’s for video streaming, Teens for Audio streaming</td>
<td>Multitasking allows streaming while doing other activities</td>
<td>2GB/day for video, 1 GB/day for audio</td>
</tr>
<tr>
<td><strong>Computing</strong></td>
<td>Information acquisition, processing, App store – auto updates</td>
<td>Browser, PC substitution, notifications, Facebook, Windows for iPad, Siri</td>
<td>95% Teens (social media) ~100% of Young Adults (Cloud processing)</td>
<td>Cloud and social media</td>
<td>7.5 GB/mo to 14 GB/mo for Teens</td>
</tr>
<tr>
<td><strong>Storing</strong></td>
<td>Cloud storage, photo and video sharing through Cloud</td>
<td>iCloud™, music, photo library, storage on Facebook</td>
<td>Teens and Young Adults</td>
<td>iCloud™, aspects of Facebook</td>
<td>&lt;1 GB/mo moving to 1 GB/mo, 5.5GB/mo extreme</td>
</tr>
<tr>
<td><strong>Gaming</strong></td>
<td>Casual with ads, low interaction to immersive networked shooters</td>
<td>Angry Birds with ads, Words with Friends, RuneScape, OnLive, 2nd Life, Farmville</td>
<td>3.4% of population generates 99% of gaming traffic from teen interactive gaming. 94% of sessions remain casual (low network) for middle age gamers.</td>
<td>Apps using widescreen and tablet formats, low latency in 4G, sensors immersing gamers in environment, increasing retirement of middle agers with more leisure time.</td>
<td>3:38 hours: min per day casual. 5.6 GB/day for future intense gamers</td>
</tr>
<tr>
<td><strong>Communicating</strong></td>
<td>Texting, voice, video, conference calling</td>
<td>SMS, iChat, Pinger, Skype,</td>
<td>Teens: 3h/day Young adults and professionals 1.8h/day</td>
<td>Persistent video via social networking. Friends are “always” within sight.</td>
<td>53MB/mo/sub to 15 GB/mo for 1 hour video/day.</td>
</tr>
</tbody>
</table>

At an extreme, we consider that if 100 percent of the mainstream population were to adopt the behavior of the leaders in each category, it would have an impact. We show this upper bound in the following figure.
Figure 19: Unconstrained growth in demand assuming 100% mainstream acceptance of leader behaviors. (This is an upper bound on the traffic forecast.)

With a more subdued adoption rate, evolving from the same 2011 starting point but only with half the adoptees in 2012, it is believed that a reasonable estimate of the unconstrained demand that will reflect the overall demand (not necessarily what will be supported), is as shown below. No caps or Wi-Fi offload are included. Notice that this shows streaming is an even more prominent application than seen in the previous mainstreaming case.

Figure 20: Crossing the Chasm estimate of the demand curve without caps or Wi-Fi offload.
With Wi-Fi offload, the traffic on the 3G and next-generation networks is reduced as shown in the following curves. The offload is different for the different applications as each app tends to be used in particular locales, some of which do not lead themselves to Wi-Fi access. Consequently, the Wi-Fi offload factor varies from about today’s 35 percent to 68 percent in 2016.

![Total USA Daily Traffic By Year](image)

**Figure 21:** US Daily Traffic forecast with and without expected Wi-Fi Offload included.

The in-home Wi-Fi offload reduces the total 2016 traffic by 31 percent while small-cell public Wi-Fi reduces HSPA/LTE traffic by another 37 percent to 32 percent of the total traffic over all Wi-Fi and 3GPP access links. With Wi-Fi offload, the growth from 2011 to 2016 is by a factor of 24 times.

### 2.4.2 MACHINE TO MACHINE (M2M) CONTRIBUTIONS TO TRAFFIC GROWTH

There has been much speculation on the potential for M2M wireless communications as an emerging new usage scenario. There has been a number of working definitions of what is meant by this classification; some have included OnStar™ and eBooks such as Kindle™ as examples of machines. However, for our consideration here, we take the CTIA definition of M2M as “applications or mobile units that use wireless networks to communicate with other machines. These applications may include telemetry and telematic devices, remote monitoring systems (e.g. smart grid, healthcare, transportation, etc.) and other devices that provide status reports to businesses’ centers (e.g. operations, traffic management, data management, etc.).”

With such devices in mind, Analysys Mason and Strategy Analytics forecasts:

---


<table>
<thead>
<tr>
<th>Year</th>
<th>Global M2M devices</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>62 million</td>
</tr>
<tr>
<td>2015</td>
<td>615 million</td>
</tr>
<tr>
<td>2020</td>
<td>&gt;2.1 billion</td>
</tr>
</tbody>
</table>

Year over year growth while impressive is declining from a bit above 52 percent to under 40 percent in 2016. About two-thirds of new M2M modules shipped in 2011 used GSM/GPRS, reflecting the low data rate demands of these applications. The M2M market tends not to need high data rate but rather good coverage and long service life.

Moreover, the growth in M2M is less than the growth rate of smart phones so the overall fraction of traffic used by M2M is actually seen to decline for the past several years and in these forecasts through 2016 and beyond.

### 2.4.3 APPLICATION TO HISTORICAL AND CURRENT DATA TRAFFIC

Such effects as replacement technologies like VoIP above, or continued growth after a transition, such as one may experience as one discovers new apps or usage scenarios.

Reports are that smartphone users grow their data usage 89 percent Year over Year (YoY), going from 230 MB/month to 435 MB as of mid-2011.

The general trend, as reported recently, has seen about 35 percent to 40 percent YoY increase in traffic use by smartphone users of a single model smartphone, but when a new smartphone is introduced there is a jump in usage as shown in the following graph.

---


Cisco’s VNI data on global mobile data traffic (Figure 23), which makes no distinction between historical and forecasted data, yields vastly different projections depending upon how much of the historical record is included in the fit. The following graphic shows the impact of selecting pre-2012 data points (in blue) and including Cisco’s own forecast to 2015. We see, then, that we live in a critical time, when the wireless data explosion is upon us, but its further outlines are only hinted at in current data. Careful analysis of the traffic statistics in the coming quarters may reveal the 2015 and 2020 expectations or perhaps see the unleashing of further pent-up demand.

---

These forecasts of overall traffic do not directly reflect the signaling traffic load. We know that LTE signaling is more efficient than WCDMA signaling so the signaling can be expected to grow somewhat more slowly than the overall traffic.
3 RECOMMENDATIONS

3.1 ADDRESSING IMPACT OF APPLICATION AND DEVICE PLATFORM BEHAVIOR

Many of the impacts on wireless networks due to applications are addressed in the sections below. The asymmetric nature of traffic in wireless networks should be addressed by using the flexibility in the RABs in HSPA, and appropriate scheduling of packets over the networks in both HSPA and LTE. Spectrum allocation in the uplink and downlink in LTE can be used to address this. These aspects are addressed below and in Section 3.3.1 below.

3.1.1 HEARTBEAT FOR ALWAYS-ON APPLICATION

As the application interface to mobile network, the device platform should not only provide connective function to application, but also coordinate with the mobile network to make heartbeat activities more network friendly and collaborative. For example, it may be possible to group heartbeats together instead of sending them individually as small burst packets, or to send them while UE is already in an active state.

The mobile network can share some collaboration and control capability from device platform to coordinate application’s heartbeat or push behaviors through APIs to applications and devices, based on:

1. heartbeat is a way for the application to realize always-on capability and let the application server know the reach-ability of the application client, and
2. the mobile network already is aware of when and how to reach the device

An appropriate range for timers in the network should be considered to address the periodicity of these heartbeats.

The mobile network can also provide APIs to developers performing various network performance impact measurements to give platform and application developers visibility into how their behaviors impact network performance. With other incentives from mobile operators, developers can adjust their applications and platform behaviors accordingly in order to minimize network impact. It is recommended that some of these approaches are standardized in the industry.

3.1.2 CONSTANT PUSH SERVICE

Similar to what was mentioned above with respect to heartbeats, it is recommended to group push notifications together to reduce signaling load on networks. In addition to this, user awareness of signaling load and the effect on battery consumption has to be increased to accept longer push service intervals, as covered in Section 3.6.1.

3.1.3 NETWORK ATTACHMENT

Device platforms should perform smart access technology selection and prevent frequent switching between different access technologies during short periods of time. Mobile networks should also collaborate with device platforms to prevent consistent network attachment retry from large quantities of devices while there is overload occurring in the network. For example, device platforms can implement a
back-off retry mechanism after receiving failure indications from network or detecting network traffic congestion itself.

3.2 ENHANCEMENTS TO NETWORK ARCHITECTURE

3GPP standards provide reference network architecture to enable mobile data connectivity and an essential suite of services. Supporting the numerous third party applications on this infrastructure requires additional network elements to enhance the standard architecture. The additional network elements and features may integrate functions across Layer 1 – 7 in both the Control and Data planes to enable a more holistic approach to network utilization when deployed inside a carrier network.

3.2.1 SERVICE DELIVERY ENVIRONMENTS

Historically, core applications on wireless networks such as voice and text messaging were developed by network infrastructure vendors and deployed and operated by the operators. However, the multitude of the new services being launched now are created and operated by application developers who do not have access to the details of the architectural and operational capabilities and limitations of the mobile networks.

Simultaneously, operators can sometimes be overwhelmed by the avalanching pace of innovation in applications both in terms of number of applications and their diverse requirements on the networks. Addressing this application-network awareness gap requires a mechanism for exposing network information to application developers and applications’ network resource requirements to operators.

For many applications, the network application platform is the exclusive interface to the mobile network. Application platforms need to incorporate a standard mechanism for applications to request capabilities of mobile networks such as network QoS and QCI level and to specify the minimum and maximum performance requirements of the application. The minimum and maximum performance requirements for the application could be determined by design analysis or via simulation testing.

Further, the application platforms should also incorporate a managed infrastructure for aggregating such network capabilities across applications. The managed infrastructure for network capabilities should be capable of accepting the resource requests from applications, negotiating the resource requests with the mobile network infrastructure and monitoring application behavior to detect any deviation from the negotiated network resources. Such application platform functions must integrate with and be complementary to the policy management infrastructure functions in mobile networks.

Complementing application access to network capabilities is achieved with support for network friendly technologies in the device platform. For example, adaptive streaming protocols for multimedia allow for automatic accommodation of network congestion ensuring an optimal QoE for all users and applications in a network. Other network-friendly technologies include the on-demand download of email and MMS attachments and the support for degradable QoE protocols for video conferencing.

Such functionality can be provided by a flexible Service Delivery Environment (SDE) layered on top of the robust connectivity delivered by the Connectivity layer. Figure 24 shows an example of one such SDE.
Figure 24: APIs are used to expose network capabilities to applications (e.g. OMA API\textsuperscript{28})

The SDE should include software APIs for exposing network assets and capabilities such as Network element and UE location, Security configurations, Connectivity layer QoS and Generic access bearer identification, similar to the GSMA OneAPI\textsuperscript{29} and OMA API. Further, a SDE could include a service logic execution engine and interfaces to QoS and policy control analytics and possibly billing systems to enable application developers to understand the affect and efficiency of their services. The SDE can be used by application developers in conjunction with application components resident on the device and in the cloud to enable development of applications that are aware of the mobile network characteristics including the Connectivity layer, such as, access network selections, data aggregation within terminals to avoid multiple connections and frequent signaling, and avoid synchronized network access between terminals, etc. In addition, the SDE will also enable on-device application platform developers to better integrate mobile network and Connectivity layer functions into the platforms for further exposure of network characteristics to applications through platform APIs. The SDE will also enable application developers to develop applications that span the Cloud and the Network infrastructure domains for enhanced functionality. Besides providing application-facing and operator infrastructure-facing APIs, the SDE will also need to enable the creation of service processing graphs for sequencing of the processing and ensure adequate security for the execution of the service logic.

\textsuperscript{28}http://www.openmobilealliance.org/API
\textsuperscript{29}http://oneapi.gsma.com/
In addition to deploying SDEs, carriers also need to promote the availability of such functionality among application developers. For instance, vendor and wireless industry developer conferences should include sessions to educate application developers on the SDEs. Further, adequate SLAs and appropriate business models should be developed around the SDEs to enable application developers to use them in commercial applications.

### 3.2.2 SUBSCRIBER AND SERVICE AWARE TRAFFIC MANAGEMENT

Understanding customer behavior and utilizing policy control could possibly be an effective mechanism to enhance network traffic management. Subscriber and Service aware traffic management includes the capability to identify type of subscriber, application and services in the traffic in access and signaling control network element, also using techniques such as content inspection to classify and direct traffic appropriately at Layer 3. Further, traffic thus identified can be individually addressed with service specific policy based on the nature of the traffic flow and nature of the payload. For instance, web browsing traffic is transactional and bursty while video traffic is long-lasting over an extended period of time. Traffic from embedded mobile devices may tolerate bigger latency.

Such insights into the nature of the classes of traffic can be used to process the traffic at Layer 4 i.e., manipulate the content to coalesce bursty sessions, recompress content payloads or throttle greedy transmission protocols in sync with network neutrality principles. Based on prevailing network traffic conditions, the traffic may also be blocked or deferred in extreme cases. A possible enhancement for traffic detection is to provide an application-specific user priority (or optimization index) over the interface between PCRF and PCEF, which would allow operator-defined optimization enforcement action to be signaled between these entries. Further, the radio link may also be driven factoring in the service characteristics for further optimization of OTA Data plane resources.

### 3.2.3 NETWORK OPTIMIZATION

3GPP Rel-10 and Rel-11 have added congestion control mechanisms in access network (for example in EUTRAN and UTRAN) to protect access and core network from overload caused by various devices and traffic types, especially the excessive signaling from large number of the devices and applications. A number of measurements, such as device access priorities, backoff timers and extended wait timers, extended access barring, per device periodic updating timers, etc., can be deployed to prevent signaling storms by devices and applications, which could bring down partial or even entire network if not design and engineered correctly.

Traffic in the Data Plane of the network yields to optimization both in the Connectivity and Services layer. Connectivity Layer optimization includes techniques such as TCP optimization and traffic shaping.

TCP optimization increases compatibility of the TCP protocol to wireless network characteristics allowing for increased throughput efficiencies.

Policy-based traffic shaping enables networks to cope with transient congestion scenarios by prioritizing network capacity for higher value traffic.

Services Layer optimization includes techniques such as bit rate shaping of multimedia traffic, transcoding of multimedia content to accommodate real-time network conditions and UE capabilities, reconstitution of web pages for efficient markup and UE compatibility and lossless compression of content with dictionary based algorithms. Given application traffic is predominantly transported over HTTP and
HTTP is a transactional protocol, further efficiencies can also be achieved by coalescing HTTP sessions using techniques such as image inlining, persistent HTTP sessions and in-network content caching.

Network optimization at the transport and application layers together enable enhanced utilization of network capacity by bridging application capabilities and requirements with Connectivity layer capabilities.

In a recent ATIS NetOp white paper study\(^3\), it is further suggested that 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. A number of optimization mechanisms besides those mentioned in this section are being considered, such as, congestion-aware fairness, network-aware scheduling of content, Load- and Policy-aware multi-RAN selection, optimizing use of wireless signaling resources, etc. (some of them will need further standards development.) Please refer to the ATIS white paper study for details.

### 3.2.4 DATA CACHING IN NETWORK

Data traffic on mobile networks has significant statistical redundancy due to relative popularity of some bandwidth-intensive content versus others, especially short form video. This relative popularity is especially amplified on mobile networks given mobile devices are immediately accessible and users tend to rapidly share viral content on social networks. The statistical redundancy can be used as a basis to cache static content at various nodes in the network to reduce retransmission or reduce repeating transmission of popular content. The caching architecture can be designed based on the size of the cell, and user demographics. Such caching has relevancy in the mobile packet core, mobile edge/RAN and client devices including small cells.

### 3.2.5 VIRTUALIZATION

Virtualization for compute, storage and networking brings economies of scale through improved efficiencies of resource utilization and has had a significant effect on the economics of services delivered over wired networks in this past decade. The incorporation of virtualization techniques both in the Internet and inside the operator infrastructure will similarly help redefine the economics of service delivery in mobile networks. Examples of use of virtualization in the Internet or cloud include offloading of compute-intensive media processing to compute farms or long term storage of content in storage farms. Examples of in-network virtualization include RAN sharing among operators, core network integration across smaller regions/countries (subject to regulations) and follow-the-moon processing, load sharing.

Follow-the-moon processing here refers to the relative abundance of compute cycles available during non-business hours in data centers and their lower cost due to the lower cost of electricity during these off-peak hours. Leveraging this lower cost compute cycles for compute-intensive processing such as video and analytics could enable capabilities that would otherwise be financially unattractive.

### 3.2.6 SERVICE-LEVEL ANALYTICS TOOLS

Besides better management of traffic generated by various services and increased application – network transparency, another potential source of significant performance improvements is the incorporation of

---

data-driven computing paradigms in mobile networks. It is common practice to use data-driven computing processes in the Service layer to monitor application and resource usage and further translate the statistics into productive monetization enablers. Data-driven computing enables quantification of operational metrics and an adaptive network design and build out strategy enhancing both the observability and controllability of networks.

Incorporating data-driven computing into mobile networks includes incorporation of operational analytics tools in network elements, system-wide performance logging and dashboard capabilities and research tools for diagnosing problems and performing what-if analysis scenarios for network planning and marketing purposes. Understanding network performance and utilization at such various levels of abstraction and in various levels of real-time or near real-time time frames will enable increased efficiency in current and future network operation.

### 3.2.7 TRAFFIC OFFLOADING

To evolve network architecture to meet the demand for high-performance, low-cost services, traffic offloading and integrating them into operators’ infrastructure offers transformative opportunities for mobile broadband network design. There are two types of offload for mobile operators: radio access network (RAN) offload and core network offload. These are targeting different parts of the network and impact users differently. Small cells bring capacity to where it is needed. However, there are some important links between them.

Given radio spectrum is the most expensive resource in a mobile network, offloading traffic from the licensed spectrum Wide Area Network (WAN) to other unlicensed spectrum networks, small cells or custom licensed spectrum traffic offloading networks enables improved utilization of the radio spectrum resource. As described in Section 1.3.2 above, offloading to end user owned Wi-Fi broadband terminations is an efficient mechanism since the entire responsibility of connectivity is borne by the end user. However, Wi-Fi offloading may remove the network operator from both the Control and Data planes of the Connectivity layer potentially impacting service continuity especially as core services such as voice move to IP on next-generation networks.

Offloading through small cells enables the network operator to retain management of the Control plane functions in the offloading scenario and optionally the management of the Data plane functions as well. With interoperability with macro cells resolved, small cells offer an attractive and flexible option to offload customer-premises radio resource utilization.

A final option is the emerging meta-mobile networks that offer WAN offloading with full mobility operating on alternate licensed spectrum alongside traditional 3G and next-generation networks. With the increased support for multiple radio bands enabled by LTE, these networks offer on-demand WAN network capacity.

The solution for tighter integration of Wi-Fi with cellular services is called “managed Wi-Fi offload” or “service provider integrated Wi-Fi”, to provide better service and retain usage in operators’ own infrastructure. With “integrated Wi-Fi”, Wi-Fi access will be fully part of the mobile network and integrated into the core. It will support session mobility, policy-driven, fully transparent to end users. EPC integration of Wi-Fi could look somewhat similar to LTE small cells, where IPSec termination will be needed for untrusted IP backhaul, possibly aggregation device at edge of the EPC is needed for aggregating the number of incoming tunnels/connections. This will make provision of a common service portfolio much easier and will allow operators to make better use of the technologies such as ANDSF (Access Network
Discovery and Selection Function) and policy to determine when and where a device should connect to Wi-Fi.

Besides Radio Access Network offload mentioned above, 3GPP also has several initiatives on core network offload in progress, such as, LIPA (Local IP Access) & SIPTO (Selected IP Traffic Offload) in R10 for IP traffic offload at closer to edge of the traffic, IFOM for ability to split across networks on the same device, etc. While LIPA is actually part of the Femtocell architecture, SIPTO is directly related to the core network offload. The concluded solution for SIPTO in macro cell network is to distribute packet gateways (L-PGW/GGSN) so that traffic is not concentrated on a small number of core nodes. This requires some changes to SGSNs, HLRs and MMEs, and likely local routing infrastructure at the distribution site, but in theory should be reasonably straightforward to implement. Figure 25 below shows an example with L-PGW deployed locally to eNB for traffic offload and policy control.

Providing visibility into the offloading options and the state of network connectivity in real-time will enable application developers to develop a better user experience at a lower cost of connectivity. A further enhancement would be the use of predictions on the availability of offloading options based on end user usage patterns. Along with the support for offloading, support for session continuity across various generic access networks will enable seamless switching between access channels. Such seamless mobility will help develop user experiences that adapt better to diverse generic access networks.

3.3 ENHANCEMENTS FOR CLIENT DEVICE APPLICATION PLATFORMS

A significant set of mobile network functionality is exposed to application developers through the mobile client device application platforms. In the case of many applications, the application platform is the exclusive interface between the application and the mobile network. Hence, use of good design practices in the application platform and the exposing of mobile network capabilities by the application platforms could significantly help in the improved utilization of the network.

3.3.1 EFFICIENT USE OF RADIO RESOURCES

The application platform typically manages the use of radio resources in client device. Hence, proper design of the radio resource utilization logic is critical to ensure a satisfactory trade-off between client device issues such as battery power consumption and network utilization.
3.3.1.1 HSPA RADIO RESOURCES

Fast Dormancy

3GPP Rel-8 Fast Dormancy feature helps to save UE battery when the radio layers in the network get information directly from the application layer when the data transmission is over. The Fast Dormancy functionality is illustrated in Figure 26. Fast Dormancy clarifies UE behavior, providing the network with information of what the UE actually wants to do, leaving the network in charge of the UE RRC state. The UE is not allowed to release RRC connection and move to idle state on its own. When the RNC receives SCRI from a UE with the cause value indicating a packet data session end, the RNC can command the UE to PCH state instead of releasing RRC connection and dropping the UE to idle state. This approach avoids unnecessary PS RAB setups and enables network to separate signaling connection failures from fast dormancy related signaling connection release indications.

The network also broadcasts an inhibit timer (T323), setting a minimum delay between two SCRI messages for Fast Dormancy a UE is allowed to send. This is to prevent a UE from sending a constant flow of SCRI messages if for some reason the network is temporarily unable to move the UE to a battery saving state. The presence of T323 in the broadcast can also be considered an indication of network support of the Fast Dormancy functionality triggering the UE to use the SCRI messages with the cause value indicating the PS data session end. 3GPP specification allows UE to send SCRI also from Cell_PCH and URA_PCH states but there should be in practice no reason for UE to do so.

Notably the 3GPP standard based Fast Dormancy was defined to be early implementable, i.e. even though the behavior and changes in the signaling are defined in the Rel-8 specifications it is possible to build these extensions to a UE and network that is not Rel-8 compatible. Such an approach enables fast introduction of a feature that has only a very limited set of modifications in the radio protocols.

Behavioral testing of client devices

A further measure to ensure efficient radio resource usage by client devices is to increase the behavioral testing for the certification of client devices on mobile networks. As client device technology matures and gets modularized, application development for mobile networks will extend beyond software developed
for client device application platforms to hardware-software solutions (i.e., M2M applications). As the diversity of such M2M applications increase, improved certification processes will be required to ensure proper operation of the devices in the field. Further, having a fully OTA configurable software platform on these M2M applications should potentially be mandatory to ensure the option of in-field upgrades to handle bugs, glitches and interoperability.

**Continuous Packet Connectivity (CPC)**

This improves the UE power consumption (battery life time), access time and air interface capacity. However, this requires mobile functionality according to 3GPP Rel-7. This feature enhances the "always-on" experience, reduces interference on control channel and increases air interface capacity. Other 3GPP Rel-7 improvements include HS and EUL FACH which allow more traffic on FACH, less signaling to cell DCH.

### 3.3.1.2 LTE RADIO RESOURCES

LTE is designed for flexibility with the following design considerations as a part of 3GPP standardization:

As described in Figure 27, OFDM is used in the downlink. This technology is based on a large number of narrowband signals or tones, each 15 kHz. Compare WCDMA where we have one single wideband carrier of 5 MHz. This makes it robust against multi-path propagation. In LTE uplink we also use OFDM but it is special form called DFTS-OFDM, Discrete Fourier Transform Spread – OFDM, otherwise known as SC-FDMA, Single carrier frequency division multiple access. This single carrier is achieved by grouping a number of 15 kHz tones to together in larger blocks. This wider band reduces the large power fluctuations you get in normal OFDM with all the 15kHz tones and this reduces the peak to average power ratio which is high with plain OFDM. Thus, SC-FDMA allows for simpler/cheaper PA in the terminal can be used that drains the battery less fast (a factor of two to three times is mentioned in different industry reports). This optimizes use of battery on the device as more data is consumed.
Advanced antenna solutions including MIMO allow for doubling or quadrupling data rates with high data reliability. Spectrum flexibility allows for efficient usage of spectrum, unlike the fixed bandwidths of 5 MHz in HSPA. Operators should leverage spectrum flexibility to re-farm GSM and HSPA spectrum, as needed, for efficient use of available spectrum.

Additional features related to spectrum aggregation and improvements on LTE are addressed in LTE-Advanced.

The above improvements in LTE help address somewhat both the high data usage from video (e.g. using 20 MHz spectrum) or the trickle data use of some M2M use cases (e.g. using 1.4 or 3 MHz spectrum).

To address latency towards real-time applications like video or voice, LTE is designed with lower latency than HSPA. As shown in Figure 28, LTE is designed to switch between only three states. The device is initially DETACHED at power up, transitions immediately at power up to the CONNECTED state to exchange data with the network. When data exchange stops for a configured period of time, the device drops to the IDLE state. When data needs to be exchanged with the network, the device transitions back to CONNECTED.

![Figure 28: UE/device states on LTE](image)

With the improved latency and data rates offered by LTE, devices using LTE are being built with enhanced memory, processing power and screen capabilities like HD etc. This leads to the corresponding load on the networks as they are expected to deliver more data at lower latency than before.

In addition to the radio interface improvements, LTE/EPC (SAE) architecture separates the signaling plane for packet data (MME) from the data plane (SGW, PGW). This allows for independent scaling of the data plane and signaling plane as different types of devices load the network.

To address signaling and data traffic explosion, some client device considerations are as follows:
QoS based scheduling of traffic: This ensures that real-time traffic like voice and video carried over LTE is appropriately prioritized.

Choice of scheduling algorithms aligned with device/subscriber distribution: This ensures that scheduling strategies for small, high capacity cells are different from that for large cells with diffuse capacity needs.

Use of network features like eMBMS that optimize use of radio capacity by broadcasting data used by a large subset of subscribers in an area.

Appropriate use of radio resource idle timers to mitigate effect of long hold times on data sessions.

Appropriate design of APNs to ensure that resources assigned to data heavy services like video are segregated in resource allocation. This allows for dropping of these services as needed in a multiservice end user offering.

3.3.2 APPLICATION PLATFORM SUPPORT FOR NETWORK CAPABILITIES

Applications run on the device application platform. For many cases, the device application platform is the exclusive interface to mobile network capabilities provided by the device. Hence, the application platform needs to provide functionality beyond basic connectivity by providing connectivity layer capabilities such as Control Plane signaling functions, Data Plane QoS functions, based on the connectivity requirements of the application.

Application platforms need to incorporate a standard mechanism for applications to request capabilities of mobile networks such as network QoS and QCI level and to specify the minimum and maximum performance requirements of the application. For minimum and maximum performance requirements for the application could be determined by design analysis or via simulation testing.

Further, the application platforms should also incorporate a managed infrastructure for aggregating such network capabilities across multiple device applications. The managed infrastructure for device platform capabilities should be capable of accepting the resource requests from applications, negotiating the resource requests with the device platform towards the network and monitoring application behavior to detect any deviation from the negotiated network resources. Such application platform functions must integrate with and be complementary to the policy management infrastructure functions in mobile networks.

Complementing application access to network capabilities is the support for network friendly technologies in the device platform. For example, adaptive streaming protocols for multimedia allow for automatic accommodation of network congestion ensuring an optimal QoE for all users and applications in a network. Other network-friendly technologies include the on-demand download of email and MMS attachments and the support for degradable QoE protocols for video conferencing.

3.3.3 APPLICATION QUALITY ASSURANCE

Improving the viability of mobile network application businesses hinges on being able to support greater numbers of end users for the applications and being able to offer improved QoE while doing so. Besides the technological enhancements targeted at improvement of E2E performance presented here, a key component of ensuring end user satisfaction is being able to offer predictable a QoE.
A predictable QoE requires numerous efforts including well designed applications, adaptability of applications to various network conditions, education of end users on the expected QoE of applications and ensuring the end-to-end security of networks. Ensuring security requires certification of applications by a trusted authority – a role that is performed by many application platform vendors. Where such certification systems do not exist, mobile network operators or third party certification houses should perhaps offer an adequate certification system. Besides ensuring security, a key secondary goal of a certification system could be being a trusted source of expected QoE of a service. Managing end user expectations could contribute significantly towards ensuring adoption of applications and hence health of the entire ecosystem. A healthy application developer community on an application platform will also have a symbiotic and salutary effect on the growth of the application platform in itself.

3.4 GUIDELINES FOR APPLICATION DEVELOPMENT

The availability of SDKs from device platforms, such as IOS, Android, Windows Mobile and Blackberry OS, and storefronts to launch applications has enabled many programming enthusiasts to come up with applications that find their use in a plethora of domains. Unfortunately, many developers in the process of providing the best possible end user experience are oblivious to the amount of signaling and data traffic their applications generate on the operator cellular network.

Smartphone applications create both signaling as well data traffic. In most cases the signaling traffic is significantly higher than the data traffic. This results in control channels getting overloaded faster than their data counterparts. Although optimizations are being done on the network side to accommodate for the influx of traffic and increase throughput, the need to design and manage smart-phone applications in a way they minimize signaling and data traffic overload on the operator cellular network has become imperative. The goal is to find the ideal configurations of mobile handsets, applications and networks to minimize negative network impact from smart devices, improve resource consumption as well as better handset and application performance and to ultimately provide the best possible user experience.

This section recommends some of the best practices a smart-phone application developer should consider while developing their application in order to create a win-win situation for end users as well as for cellular network operators. The recommendations in the following sections are limited to the four popular platforms; namely Android™, Windows Phone™, iOS™ and BlackBerry™. It is important to note that not all recommendations outlined here may be applicable to all network dependent smartphone applications.

3.4.1 ASYNCHRONOUS REQUEST AND RESPONSES:

In order for an application that fetches content from a remote server and renders results to the user, an important consideration to keep in mind is that the cellular network can introduce latencies in delivering the content or the content can get lost without being delivered. In some cases the request might not reach the server. In order to provide a smooth user experience the application may send parallel requests for the items in a page and provide an indication to the user that the content is being fetched. For example, a webpage could have thumbnails or placeholders for images it has to fetch from the server. The application must be able to process the response for an item it has received from the server and dynamically update the thumbnails or placeholders as and when the content is delivered. This would give the user the ability to take an action on the already delivered content without having to waiting for the all

---

31 Nokia Siemens Networks Smart Labs - Understanding Smartphone Behavior in the Network
the items to be rendered. Another advantage of sending parallel requests for the items in a page is that the application may implement a timeout period to receive the response for an item. If the application fails to receive a response for the item within the timeout period it can close out the data connection and free up network resource. This would also help reduce the battery consumption of the mobile device.

3.4.2 SAVING DATA AND SESSION INFORMATION:

Application developers when designing a network dependent application have to account for the unreliability of the cellular data network and that their application can encounter loss of data network at any time. Applications shall monitor the connection status and register with the device platform for network state changes. The application must save any relevant information about the current session that would help it resume the session in case of a network connectivity failure. Progressively saving downloaded data in a temporary downloads folder would also help resume the data download from the point at which it was interrupted without having to download the entire content again. This would result in bandwidth savings for the operator and for the end user as well.

3.4.3 CACHING DATA ON THE DEVICE CLIENT APPLICATION AND ON THE SERVER

Caching of data becomes essential in a number of ways: (a) Reduce the amount of signaling between the handset and network; (b) reduces the amount of data that needs to be transferred; (c) increases the speed of displaying results thus enhancing customer experience.

Caching can be implemented on the handset application and on the remote server. Figure 29 shows a typical client-server application model. Before sending the request to the server for data the application checks its local cache for the availability of the data and its validity. If the available data is valid the application presents the data immediately to the users without requiring any network interaction. If the data needs to be validated, in case the application wants to ensure that it has the latest version of the content, the application can include a checksum or last modified date of the content in its request to the server. The server compares the checksum or the last modified date and if it does not find a newer content it notifies the client that the no new version of the content is available. The client can thus render the data immediately to the user without requiring downloading the content from the network. For example, in SYNCML DS the client and the server exchange sync anchors to determine whether the content is up-to-date or needs to be refreshed.

It is important to note that not all data can be cached due to security reasons or due to the nature of the data. For example, streaming data cannot be cached due to its real-time nature. It is also important to take into consideration the memory limitation of the device before defining cache size. It may be worthwhile to let the user decide the cache size.
3.4.4 NETWORK TECHNOLOGY AWARENESS

Smartphone applications must be aware of the network technology the device is connected to (Wi-Fi, 3G/4G, 2G). They must dynamically adjust their behavior and response depending on the underlying network technology. Most operating systems publish information about the network state and applications must register with the system to retrieve this information. Information about signal strength, bandwidth and latency can be taken advantage of before sending or requesting data from the Internet. For example, a media streaming application can request a lower resolution video from the media server if a device is connected to a 3G network, or if the device is connected to a slower 2G network the application can close out the streaming session, inform the user of the slower connection, and advise the user to restart the streaming when the device is connected to Wi-Fi or next-generation network. This might also result in freeing up valuable dedicated channel resources and greatly conserving the consumption of device battery. Another example where knowledge of network technology can become useful in conserving device battery is when a device is trying to upload a 6 MB song\textsuperscript{32}. Over the 2G network the device would consume approximately 45 mAh of battery power, 9.5 mAh on a 3G network, and 4.4 mAh on a Wi-Fi network, to upload the song. Applications that upload media such as photos and videos automatically to web server can take advantage on the network technology state and defer the upload of the media until the device is connected to Wi-Fi or next-generation network. Bandwidth intensive applications shall also configure the default technology for data access to be Wi-Fi, which is generally considered as faster and cheaper as compared to the cellular data network.

3.4.5 LENGTHENING DATA BURSTS

A data burst is defined as a complete data transfer of any size. Sending data in small bursts for applications such as audio or video streaming requires the device to constantly change radio resource state from idle to the full power state (Cell_DCH state). This requires about two seconds to get done, introducing additional delay in transferring content, and draining of the device battery. Large burst are device energy and network efficient as the device does not have to keep changing states resulting in fewer control messages to change states and reduced latency. For example, to send a data packet that is 10KB large in chunks of 2KB each would take approximately 80 seconds as compared to the 16 seconds

that would take sending the same 10KB in a single burst. Researchers have analyzed\textsuperscript{33} the Pandora® Internet radio app and concluded that bundling data packets in to larger bursts results in up to 40 percent in energy savings.

![Figure 30: Comparison of signaling, time taken for small bursts and large bursts](image)

### 3.4.6 QUEUEING UP DATA

Data traffic can be categorized into two types, (i) User initiated – An action taken by the user on the Smartphone that results in data traffic and, (ii) Background data – Data traffic that could be generated after the completion of the user initiated request. For example, customer analytics information and usage statistics sent to a server. User initiated traffic is unavoidable and data has to be presented to the user immediately when requested. Optimization can be done for sending background data by looking for the availability of dedicated channel state, which could have been made possible by a user initiated request, and sending data in that time period. Thus applications can queue up their background data, which in most cases, are not time critical as they are not directly used by the end user, and defer the sending of the data to the time when the device transitions to the dedicated channel state. One other option would be sending background data when the device is connected to a Wi-Fi network. These optimizations would result in (a) Reducing the amount of network signaling to establish a data session and radio bearers and (b) Reducing battery consumption of mobile device.

### 3.4.7 RETRY SCHEDULING

The cellular wireless network is not always reliable and data packets can get lost due variety of reasons, for example, handovers. When data is lost application must check with the OS whether the device has data connection before retrying the request again. This will ensure that the retry may be successful. It would also be beneficial to spread out the retries to increase the chance of success and reduce the number of retries. This would help reduce the amount of signaling traffic generated for each retry. It might also be worthwhile to stop retrying after certain number of times and changing to a manual user initiated retry.

\footnote{33 A Call for More Energy-Efficient Apps \url{http://www.research.att.com/articles/featured_stories/2011_03/201102_Energy_efficient}}
3.4.8 TEXT COMPRESSION AND MEDIA TRANSCODING

Most web server and device platforms support text compression schemas such as GZIP. Applications must make use of GZIP to compress text format like XML and JSON. XML text of 17.5KB, and a JSON text of 13.5KB, can be compressed to just 3KB using GZIP. This would significantly reduce the amount of data transferred over the network contributing to bandwidth savings. Compressing media formats – pictures, audio and video – is not an easy task since they carry a quality element with them. For uploading pictures to a photo server such as Flickr® or Picasa®, the default size must be set to a postcard size or if the picture is being uploaded to a social networking site, which will reduce the image size anyway, there is no point in uploading the original size image. This is true for downloading images to the mobile device as well where the display size is limited and there is no added advantage in downloading the original full size image. Streaming video applications must check the display size of the mobile device before streaming content. Also the network technology to which the device is connected to must be checked and appropriate bit rate for the video must be chosen.

3.5 ENHANCEMENTS WITH RESPECT TO SECURITY

Given the aforementioned potential for security threats with respect to next-generation mobile networks, especially given the ever-increasing innovative capability of smartphones, it is imperative that appropriate measures be taken to mitigate such threats. The following provides a few considerations in this regard.

3.5.1 MALWARE DEFENSES

The first basic step in defending against malware is to deploy anti-virus and firewall software on all devices accessing the mobile network. In the wired broadband world, many operators offer such software for free to their subscribers, others for an incremental fee.

Mobile operators have themselves begun rolling out managed security services and/or network-based products with some built-in security. For example, some operators may deploy network-based security services, which protect mobile devices from viruses, worms and Trojans that can infect devices and spread malware via text messages and/or Bluetooth connectivity. This network-based service will also block Denial of Service attacks and restrict network traffic based on source, destination, IP ports and applications. It will also allow enterprise IT managers to lock and/or delete data on lost or stolen devices.

Mobile operators should also consider deploying content scanning technology in the network. For example, operators who allow a wide range of content-multimedia messages, music and video download, mobile Internet surfing and mobile blogging, they should include a content security module which will examine all elements of message content and then filter out harmful content such as viruses and spam.
It is imperative that mobile network operators protect their networks at multiple levels:

- **Packet level**: through prevention methods provided by stateless firewalls, which determine whether a packet is permitted into the network by analyzing basic information in packet headers

- **Session level**: through prevention methods provided by stateful inspection firewalls, which monitor and control the flow of traffic between networks by tracking the state of sessions and dropping packets which are not part of authorized sessions

- **Application level**: through prevention methods provided by Intrusion Detection and Prevention systems (IDP), which monitor and analyze network traffic for signs of attacks

Firewalls can help operators control fraudulent activities, mitigate threats from hackers and provide added visibility into network operations. In the case of mobile operators using GPRS/UMTS and/or LTE for example, they would require firewall products, which provide infrastructure protection from attacks across all the main interfaces with external and internal networks interfaces. It is recommended that this particular product also include VPN, packet filtering and traffic management features to help protect mobile networks from Internet borne attacks such as DoS. Another desirable feature would be, for example, that it uses rate limiting to control the rate of GTP signaling and user plane messages so that packet core networks are less likely to be overwhelmed in a DoS attack. Obviously, the deployed firewall
also needs to be robust enough to handle the traffic, which flows through it so that the user’s experience is not negatively affected.

Intrusion detection and prevention systems complement the role of firewalls in a mobile operator’s network. IDP systems are designed to detect the presence of attacks within the traffic that is permitted to flow into the network. IDP systems perform this function by using:

- **Stateful signatures**: which track the state of the connection/traffic and scan for attacks based on known patterns
- **Protocol anomaly detection**: which identifies attacks that are masked by legitimate protocol use
- **Backdoor detection**: which is capable of detecting traffic caused by worms or Trojans
- **Traffic anomaly detection**: which compare incoming traffic volume to baseline norms so as to identify attacks that might span multiple connections
- **Network honey pots**: which serve as computer systems on the Internet which are expressly set up to attract and “trap” people who attempt to penetrate other people’s computer systems
- **Compound signatures**: which combine stateful signatures to identify attacks that might span multiple sessions

IDP is usually placed behind the firewall so that the device can inspect the packets entering and exiting the network. Should malicious traffic be detected, the IDP device will sever the connection so that it never enters or leaves the network. Having IDP in place on the egress link is important because it allows an operator to prevent attacks originating within its network from impacting other operators.

### 3.5.2 PACKET FILTERING

To protect networks from bad packet injection attacks, one should provide some type of filtering mechanism to certain network elements such as GGSN/PGW to filter out such packets.

![Figure 32: Packet Filtering using Access Control Lists](image-url)
Packet filtering using Access Control Lists (ACL) should be provided for mobile network elements such as GGSN/PGW in order to control the inbound and outbound traffic to/from the mobile network.

Some filtering rules which can be applied are as follows:

1. Source Routing should be disabled.
2. Source IP addresses from un-trusted network should be filtered. Packet is a defense against IP spoofing attacks. The gateway to a network usually performs ingress filtering, which is blocking of packets from outside the address of an internal machine. Ideally, the gateway would also perform egress filtering on outgoing packets, which involves the blocking of those packets using a source address, which is not topologically correct with the current mobile network. This prevents an attacker who has inserted himself within the mobile network from launching IP spoofing attacks against machines existing both inside and outside the mobile network.
3. Subnet-directed broadcasts from un-trusted hosts should be disabled.
4. The following ICMP-Messages coming from the external network should be denied:
   - echo-request
   - redirect
   - address-mask-request
5. Only the following ICMP-Types coming from the internal network should be permitted:
   - echo-request
   - parameter problem
   - packet-too-big
   - source-quench
3.6 END USER BEHAVIOR

Improvements to network, device platform, and application design can all help alleviate capacity issues, but there can be no lasting success without understanding end user behavior.

There is no question that end users are using more and more data every year. In 2011, 44 percent of mobile subscribers owned smartphones, which is more than double the market penetration of two years before. Between 2010 and 2011, the numbers of smartphone subscribers engaged in various rich media activities other than mobile video have all grown by 45 percent or more. Numbers of subscribers performing game downloads and playing online games both increased more than 80 percent each in the same year. (Nielsen, 2011).

Through leverage of education, the right sorts of incentives, and automated assistance tools, software and platform developers can reach this wide audience in new, innovative, and hopefully effective ways.

3.6.1 INCREASING AWARENESS

Consumer awareness is the first key to enacting behavioral and habitual changes that will eventually reduce data usage. The simplicity and general appeal of on-device application has been very successful at increasing popular awareness and usage of data plans. However, with recent trends toward rapid growth in mobile data usage, it is imperative that users be educated about the effects of their personal data use and how to minimize their usage footprints.

There are many current efforts to increase user awareness of their personal data usage. Mobile operators have long provided network-based usage meters and limited data plans to try to raise awareness of usage and provide disincentives to over-use. Some operators have introduced data caps and throttling thresholds, and offer notifications of approaching limits via SMS and/or email. One major device operating system has recently incorporated a data usage monitor, allowing users to set an on-device notification threshold as well as monitor the data usage of individual applications. This approach affords greater awareness of total data usage to users, as many operator-provided data counters will usually exclude unbilled usage; such as the communications of operator-provided applications, or those to and from white-listed IP addresses and/or URLs.

An example of behavior that could possibly benefit from industry focus is the often misused technology of “Push” email. This technology affords rapid delivery of email to a mobile device once it hits the recipient’s email server, which appeals to users who want to feel constantly connected and better informed. For users who do not require that level of connectivity, Push email actually functions less effectively in a wireless environment. In Push solutions employing a persistent HTTP session (e.g. Microsoft® Exchange ActiveSync®), normal network interference, connection drops, etc. all result in more overhead than theoretically envisioned by the Push solution architects. In Push solutions employing proprietary network components (e.g. BlackBerry®), the extra overhead is not experienced by the end user.

In addition to improving the current efforts to educate users, there are further opportunities for influencing end user behavior. Among these include the chance to educate users about the convergence of voice and data. If users can be encouraged to monitor their data usage the way they have learned to monitor their voice minutes, they will be better equipped to deal with any future changes in service delivery, billing, and overall functionality.

There is also a strong opportunity to encourage consumers to have a greater world-view in mind as they use their devices. Early adopters of the new generation of smart phones experienced capacity issues
first-hand when their operators saw unanticipated Petabytes of additional monthly traffic generated by the new devices (based on an example of ~400 MB/user/month and ~7M user estimates; Cauley, 2009). What if those users had been able to access data on the overall effect of their usage patterns? What if their local tower reported how close to connection capacity it was? How about if users had been given information about when network “prime” times are, much the way a state Department of Transportation might periodically list the traffic levels for commuter traffic, bridges, toll roads, etc.? Users educated about their mobile data ecosystem and given tips for being a good “neighbor” in that ecosystem will have the greatest opportunity to enact widespread change.

3.6.2 INCENTIVES FOR REDUCING DATA USE

With proper awareness of the amount of data they use and the effects of over-use, many consumers may still require extra incentives to modify their data habits. Currently there are several efforts to discourage data over-use by some operators, including deployment of data caps, throttling thresholds, and assessment of overage charges.

There may be opportunities to take advantage of current trends as other incentives for reducing data use. To ensure efficiency of mobile apps, third-party developers who employ “data conscious” usage profiles could possibly be rewarded for their efforts. There are likely many other opportunities to provide incentives for “good behavior” in the mobile data ecosystem – we as an industry simply need to find and capitalize on them.

3.6.3 EQUIPPING USERS FOR EFFECTIVE USAGE MANAGEMENT

Once consumers are aware of their personal usage and have the proper incentives to manage their bandwidth, they will look to developers to provide them with effective tools to manage their data use and adapt it to their usual daily needs. Current management tools include customizable sync schedules to allow users to set the minimum frequency of updates they need. Likewise, sync detail levels can also be customized – with most business solutions, users can choose to sync only email header information, whether to include attachments or not, or even just to sync contacts and calendar. Operator applications and network counters will often inform a user of surpassed thresholds and impending limits to their data use. Some operating systems have incorporated simple but useful utilities to monitor total data usage as well as usage on a per-app basis, additionally allowing customizable warning thresholds and upper limits.

A good example of an inclusive tool kit might incorporate widgets displaying total usage, apps currently using data, and a list of watched apps and their data usage. It allows you to set alerts and warnings, about apps that are using a lot of data, approaching data caps, or when roaming. It could include a feature to predict when you will reach your monthly cap based on month-to-date usage. Users might also be allowed to restrict specific apps to Wi-Fi, or block data once a data cap has been exceeded. Finally, it could incorporate advice from other users and tips from operators about how to maximize data and which plans have the best value based on actual usage patterns.

Even with the existing tools, there are still opportunities to assist users with managing data usage. Some modern smart phone platforms still lack integrated or third party solutions that provide robust and feature-rich data monitoring and managing features. Users might eventually be able to set their own throttle speeds on a per-app basis with improvements to network QoS mechanisms (see section 1.5 above). Finally, users may be given tools and encouragement to set their own hourly/daily/weekly usage caps or schedule peak and idle time use-patterns.
4 CONCLUSIONS

This white paper identified the changes required in wireless broadband ecosystem today due to the explosion of devices, applications and increased user expectations on data performance over wireless. This includes recommendations on wireless infrastructure, devices, mobile OS, and also on user behavior.

Mobile Broadband capacity needs to continue to increase with new technologies and powerful devices entering the market. Novel applications on the new devices in the market, especially video applications, are driving exponential growth in data. Bursty and chatty applications like Facebook and enterprise applications are increasing signaling traffic, leading to signaling and data congestion. Challenges in the device OS and application space are contributing to network load and need to be addressed.

The rapid growth of wireless data and traffic has created a need for additional spectrum. The industry is challenged to use available limited spectrum in an efficient manner. Efficient use of radio resources and ideas on use of the best available access could be considered to mitigate the effects on the cellular networks. Some of these strategies include traffic offloading to the best available access or alternate wireless technologies like Wi-Fi. Network architecture evolutions like small cells and femto networks could be considered.

The use of Quality of Service to assure end user experience is getting to be more important than best effort traffic in congested mobile networks. Collaboration between network providers, application providers and device manufacturers is critical to address capacity needs and end user expectations. Improvements in device, application design and network feature activations are needed. Security concerns on the fixed Internet are now moving into the mobile broadband ecosystem. Changes are needed on mobile security infrastructure to address new vulnerabilities introduced into application driven mobile broadband networks.

Smart usage of network resources is needed to address the predicted growth in data and signaling traffic. This will require collaboration between application developers (using the device resources judiciously), device manufacturers (leveraging radio and device capacity conserving features efficiently), network operators (optimizing networks to handle varying traffic needs) and network vendors (driving efficiencies into network equipment).
5 REFERENCES

4G LTE / LTE Advanced for Mobile Broadband – Erik Dahlman, Stefan Parkvall, Johan Skold

Nokia Siemens Networks Smart Labs “Understanding Smartphone Behavior in the Network”
www.nokiasiemensnetworks.com/smartlabs
The mission of 4G Americas is to promote, facilitate and advocate for the deployment and adoption of the 3GPP family of technologies throughout the Americas. 4G Americas’ Board of Governor members include Alcatel-Lucent, América Móvil, AT&T, Cable & Wireless, CommScope, Entel, Ericsson, Gemalto, HP, Huawei, Nokia Siemens Networks, Openwave, Powerwave, Qualcomm, Research In Motion (RIM), Rogers, T-Mobile USA and Telefónica.

4G Americas would like to recognize the project leadership and important contributions of Gautam Talagery of Ericsson, as well as representatives from the other member companies on 4G Americas’ Board of Governors who participated in the development of this white paper.