University of Würzburg
Institute of Computer Science
Research Report Series

Research Report On Signaling Load and Tunnel Management in a 3G Core Network

Florian Metzger, Steffen Gebert, Katharina Salzlechner, Albert Ratetseder, Peter Romirer and Kurt Tutschku

Report No. 484 September 2012

[1] Chair of Future Communication
University of Vienna, Austria
florian.metzger@univie.ac.at

[2] University of Wuerzburg, Germany
Institute of Computer Science
Chair of Communication Networks
Würzburg, Germany
steffen.gebert@informatik.uni-wuerzburg.de

[3] FTW - Forschungszentrum Telekommunikation Wien, Austria
romirer@ftw.at

Research Report On Signaling Load and Tunnel Management in a 3G Core Network

Florian Metzger, Katharina Salzlechner, Albert Rafetseder, Kurt Tutschku
Chair of Future Communication
University of Vienna, Austria
florian.metzger@univie.ac.at

Steffen Gebert
University of Wuerzburg, Germany
Institute of Computer Science
Chair of Communication Networks
Würzburg, Germany
steffen.gebert@informatik.uni-wuerzburg.de

Peter Romirer
FTW - Forschungszentrum Telekommunikation Wien, Austria
romirer@ftw.at

Abstract

This paper takes an initial look on the signaling behavior of mobile devices in mobile cellular core networks. In contrast to most contributions in this field, our focus does not lie on the wireless or user-oriented parts of the network, but on signaling in the core network. We evaluate GTP Tunnel Management signaling messages on the path between SGSN and GGSN, and the overhead imposed on user plane traffic. Furthermore, we attempt to correlate PDP Context durations to device specific parameters such as device class or operating system. From the context durations, we estimate the signaling overhead devices burden on the core network.

1 Introduction

Given its roots as a research network and its growing pervasiveness over the last forty years, it is understandable that research on the Internet covers all parts of the network from applications to access to the core, and has been going on ever since what could be considered prehistoric times of the Net. The state of research on mobile cellular networks such as 3G is lean in comparison. Mobile networks providing Internet access have not been around for too long, and still are not available in all parts of the world. Furthermore, most research focuses on user-oriented metrics such as traffic statistics and mobility patterns, or takes into account the radio part of the network only. Little has been published about activity within the core network, and yet less about signaling.

Given how limited spectral resources on the radio interface are, it might not seem obvious to think about signaling load in the network. Yet, there have been situations where the core network unintentionally has been flooded with signaling, taking down user-plane connectivity on the way, despite small amounts of actual user traffic being transported [1, 2].

The adverse effects of state-keeping in network devices have been known to, e.g., Internet users running BitTorrent across low-end home routers as of the early 2000s. In Universal Mobile Telecommunications System (UMTS) mobile networks, the networking hardware is vastly more powerful, but the control plane tasks are vastly more complex than port and network translation as well, namely carrying and routing IP and voice traffic, user mobility, Authentication, Authorization and Accounting (AAA) and so on. Many specialized protocols are involved to communicate intents and states in the network. This causes processing overhead, additional traffic on network paths, and increases the number of states to be held in memory on the core network nodes. Therefore, in scenarios such as the ones mentioned above, radio access is not the bottleneck to connectivity any more, but signaling is.

The inherent complexity of signaling in mobile cellular networks is easily missed by programmers who do not or cannot know that their applications will run over such wireless links, and probably would not expect it from a network that pretends to transparently carry IP. What furthers this problem is the lack of literature on the theoretical and practical sides of these issues.

This apparent lack is due to a number of reasons. First, gaining sufficiently intimate knowledge on the huge corpus of Third Generation Partnership Project (3GPP) Technical Specifications is a laborious task. Second, to come up with lower-layer measurements requires physical access to the core network infrastructure and suitable measurement equipment. Also, much of the data is commercially and privacy-sensitive, and cannot be published without extensive sanitizing.

The purpose of this report will therefore be to give a 3G tunnel management primer, introducing the relevant GPRS/UMTS network structure and the involved control plane protocols with a special focus on the GPRS Tunneling Protocol (GTP), which is probably the most prevalent. Furthermore, we share our first insights into one practical aspect of the signaling process, the GTP tunnel management procedures. Using a week long mobile network data set recorded at the Gn interface between the Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN), we take a look at Packet Data Protocol (PDP) Context durations, i.e. the time a PDP Context is established and held, argue how this influences the load on the network, and evaluate the data by device types and operating systems.

Our measurement data backs up a number of straightforward assumptions on the behavior of different device and operating system types, but also reveals some remarkable differences in tunnel characteristics.

The rest of the report is structured as follows. Section 2 discusses relevant work in the field. Section 3 introduces UMTS as well as GTP basics and protocol details relevant to core signaling. Chapter 4 gives an overview on the METAWIN data acquisition platform. Chapter 5 evaluates our data set, and Chapter 6 concludes the report and gives an outlook to future endeavors.

2 Related Work

Recently, stories about signaling storms and overloaded control planes in mobile networks reached popular news media [2, 1]. These stories blame a specific combination mobile device type, operating system and application to cause excessive amounts of signaling in the network. The Android version of the popular casual game "Angry Birds" is a free download, and uses regularly refreshed advertisements to achieve some form of financial compensation for the authors. Now imagine a large amount of devices setting up and tearing down data connections only to retrieve new ads and therefore causing tens of control plane messages on each retrieval, which could strain the signaling-heavy structure of current networks.

The dynamics behind such events are worth investigating, and some work has already been done by several publications. While these touch parts of the areas tackled in this report to some degree, we think that the combination of the focus on core signaling, PDP Context durations, and investigating the influence of devices on these are genuine contributions of our work.

When control plane aspects of mobile networks are considered, the investigation usually focusses on the radio interface and Radio Resource Control (RRC) signaling, but pays little attention to aspects in the core network. A report on cross-layer interaction in mobile cellular networks falls into this category [3], discussing interaction, e.g., between the application layer and the RRC (such as seen in the "Angry Birds" case) and its consequences for device energy consumption and radio channel allocation efficiency. The authors argue that there is much room for improvement in this area, and propose some enhancements.

In [4], mobile network traces are used to simulate a malicious signaling storm by transmitting low-volume user plane traffic with inter-departure times slightly larger than the transition timers in the RRC state machines. This constantly causes signaling to occur. The authors propose tools to detect this, and discuss possible scales of this type of denial-of-service attack. In [5] Ricciato et al. also take the approach of describing attack surfaces of cellular networks, especially denial-of-service attack made possible due to the complexity of the control plane.

Recent publications concerning device differentiation in mobile networks usually either focus on the user traffic dynamics [6], or on mobility and the temporal and spatial variations of user traffic resource usage [7].

In 2006, Svoboda et al. [8] conducted a core network measurement study of various user traffic related patterns, and also provided an initial insight into PDP context activity and durations. Finally, a recent publication [9] provides an investigation probably closest to our approach, however again aimed at RRC signaling on the Iu-PS link and not at GTP signaling at the Gn path (both of which somewhat intertwined however). The authors classify their evaluations based on device model and vendor and on the application type, and find that different devices have strongly different RRC characteristics, which could possibly also have an impact on GTP signaling.

3 GPRS and Tunnel Management

This section starts with a primer on cellular data network basics, and then moves on to describe relevant details of GPRS Tunneling Protocol (GTP), the tunneling protocol under investigation.

3.1 GPRS Fundamentals

Before diving into specifics of GTP messaging, we give a short overview on the packet switched domain of an Universal Mobile Telecommunications System (UMTS) network. This domain is closely related to the General Packet Radio System (GPRS) part introduced for GSM. UMTS, first defined by the Third Generation Partnership Project (3GPP) in Release 99, focuses its improvements over Global System for Mobile Communications (GSM) mostly on the radio aspects, while keeping the core network GPRS architecture intact at large. 3GPP Technical Specification (TS) 23.060 [10] defines the basic aspects involving GPRS protocols and its system architecture. TS 29.060 [11] describes the specifics of GTP flowing across the Gn and Gp interfaces which forms the basis for our work.

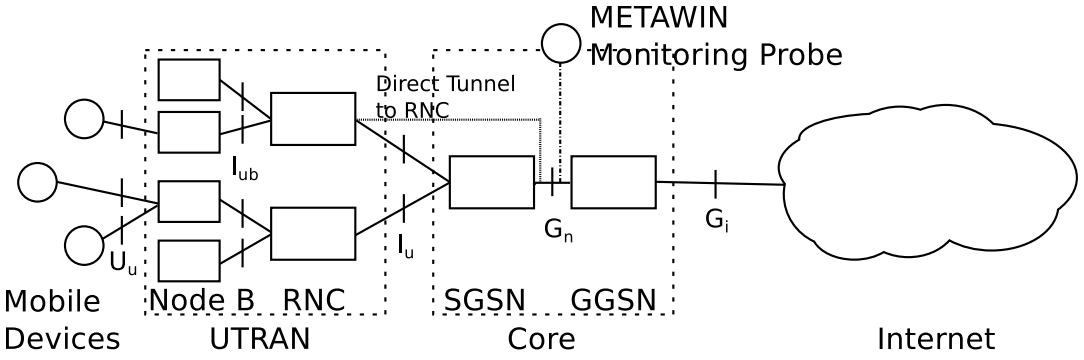


Figure 1: Typical simplified setup of the packet switched domain in an UMTS network including a METAWIN monitoring probe.

As shown in Figure 1, user traffic originating at any Mobile Station (MS) connected to the radio network flows through one of the Node Bs, providing radio connectivity. Multiple Node Bs are aggregated by a Radio Network Controller (RNC). Node Bs and RNCs form the UMTS Terrestrial Radio Access Network (UTRAN), which is typically connected by back-haul fiber links to the core network part formed by the Serving GPRS Support Node (SGSN) and the Gateway GPRS Support Node (GGSN).

One role of the SGSN is as the mobility anchor for mobile devices, and it is the endpoint for Radio Resource Control (RRC)-based signaling and the Radio Access Bearer (RAB). The GGSN provides the gateway to the public Internet. The Gn interface connects those two nodes, using the GTP protocol to exchange user as well as control plane traffic as seen in the protocol stack in Figure 2. GTP is further separated into GTP-C, facilitating control message exchange, and GTP-U for transporting user traffic through tunnels.

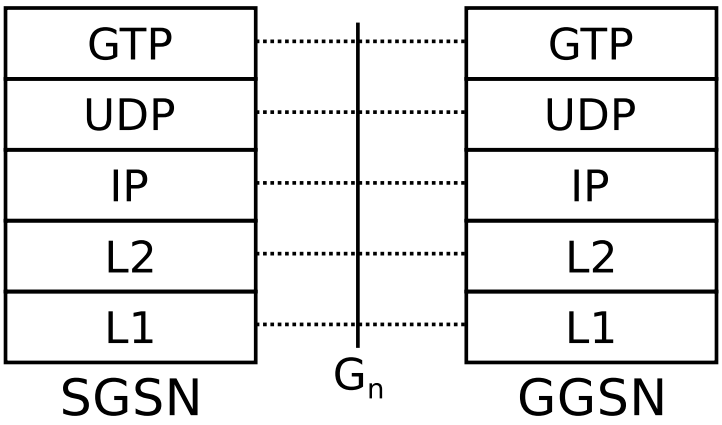


Figure 2: Typical signaling protocol stack at the Gn interface between SGSN and GGSN.

3.2 GTP Signaling

Tunnels are defined in the SGSN and GGSN in Packet Data Protocol (PDP) Context data structures. These hold various information related to a tunnel, such as the device IP address, International Mobile Subscriber Identity (IMSI), and a tunnel identifier. A tunneling concept is used for user traffic to isolate it from core network control plane traffic and to provide certain Quality of Service (QoS) guarantees to the user traffic. To distinguish multiple QoS profiles per device, up to ten additional secondary contexts can be established beyond the primary PDP context, all with different QoS allocations. However, secondary contexts are rarely in use today, and any user-plane IP traffic is transported within the primary "best effort" tunnel.

As already mentioned, GTP-C signaling is used to administer these contexts. Across the Gn path, it contains procedures for managing data paths, MS locations, mobility, and, of course, tunnels. We take a specific look at the last one. GTP messages usually come as request-response pairs. Neither part has fixed size, but is rather constructed from a number of Information Elements (IEs) of partially variable length.

The focus of our work will be the three Tunnel Management message pairs involved in the maintenance of PDP Contexts. These are the *Create, Update,* and *Delete PDP Context Requests* and *Responses*. Each pair, including their causes and possible effects, will be treated in a separate section, with the Create and Delete messages forming the substrate for our investigations presented in this report.

The variable-length nature of these messages makes evaluating the imposed network signaling load rather difficult. For example, the Create Context Response consists of up to 36 IEs, some of them mandatory, most either conditional or optional. Including the headers of both the packet and the individual elements, the minimum size (counting only the required bytes of variable length elements) is 52 bytes, while the minimal maximum size with all IEs present is 307 bytes.

Taking this maximum value we arrive at a naive estimate of the maximum overhead on user traffic imposed by tunnel management signaling in our dataset. The ratio of (tunnel management) signaling traffic to total user plane traffic is a minute 0.10%. Therefore,

the sheer volume of control plane traffic appears to be non-critical in this setup. We assume thus that the overload problems mentioned above arise rather in areas affected by signaling except for the pure transport of data, such as the memory profile of the states kept in the gateway nodes, the time required to process the large number of information held in the messages, or the imposed latency through several message round trips during transactions. The detailed mechanics of system load could be a field of investigation for future work.

3.2.1 Create Context Messages

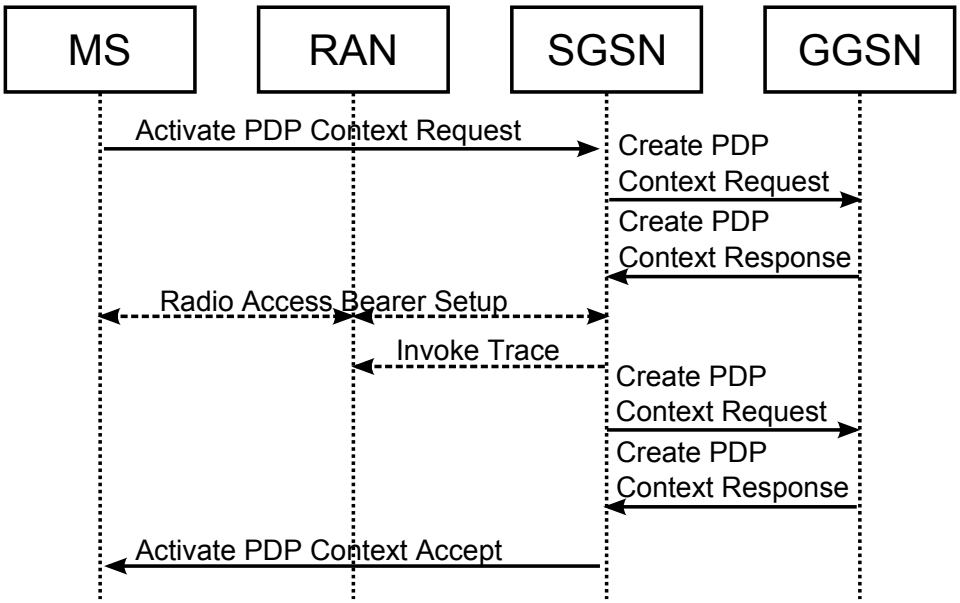


Figure 3: PDP Context Activation Procedure in a UMTS network.

Figure 3 shows the *PDP Context Activation Procedure* as defined in [10]. Some additional Customised Applications for Mobile Networks Enhanced Logic (CAMEL) procedures may be involved in the creation but are not of interest in the context of this report as they are not required. Tunneling messages are usually triggered by other procedures on different interfaces. In the case depicted here, the procedure is initiated by the mobile device through a Radio Access Network Application Part (RANAP) protocol *Activate PDP Context Request* typically sent when establishing a mobile data connection.

Generally speaking, any Create Request is part of a *GPRS PDP Context Activation* procedure, which can happen under several circumstances, the aforementioned one being the typical, but also during each *Secondary PDP context Activation* procedure for every tunnel beyond the first. When a GGSN receives this request from an SGSN, it attempts to complete the Context creation. Depending on the outcome, a response is sent back, indicating the success or failure of the operation. Typical failure codes observed in our measurements were either due to incorrect information supplied by the device ("user authentication failed"), due to malformed messages (e.g. "invalid message format"), or

indicated problems or temporary overload in the network ("no resources available" and "system failure").

3.2.2 Update Context Messages

The possible causes for an *Update Context Request* are as following.

- The mobile devices moves between SGSNs, causing a *GPRS inter-SGSN Routing Area Update* procedure.
- Parameters belonging to the context such as the assigned QoS are altered using the the *PDP Context Modification*.
- As part of *Context redistribution and load balancing* procedures.
- The MS switches between UMTS and GPRS access technologies, causing a *Inter-system intra-SGSN Update* procedure. Note that the same tunnel can be used regardless of the radio technology.
- As part of a direct RNC to GGSN GTP-U tunnel activation procedure, thereby circumventing the SGSN. Or, finally,
- To activate secondary PDP contexts using the *Secondary PDP Context Activation* as previously described.

By observing Update Context message one could, for example, capture most forms of mobility happening in the network, and get a good picture of correlations between mobility and tunneling characteristics. Additionally, tunnels using UMTS and GPRS radio technology can be distinguished, which should in theory lead to wholly different pictures, as nowadays GSM/GPRS is either used in older models or feature phones, or in mobile scenarios in rural areas where the larger GSM cells are more prevalent. Both could indicate that the data session will be rather short due to either clumsy devices or the low throughput rates of GPRS.

3.2.3 Delete Context Messages

The third type of Tunnel Management messages are the *Delete Context Request* and *Response*, indicating the immediate release of the Context involved. They are part of

- The *GPRS Detach* procedure from the SGSN to the GGSN, when a device completely deactivates its data services.
- The *GPRS PDP Context Deactivation* procedure from the SGSN to the GGSN, if only one specific tunnel is to be removed.
- The *part of PDP Context Deactivation Initiated by GGSN* procedure signaled to the SGSN.

3.2.4 Mobility and Radio-related State Machines

As indicated before, most nodes in a cellular mobile network keep all sorts of states characterizing the data connection. For the tunnel management aspects, two state machines are of special note.

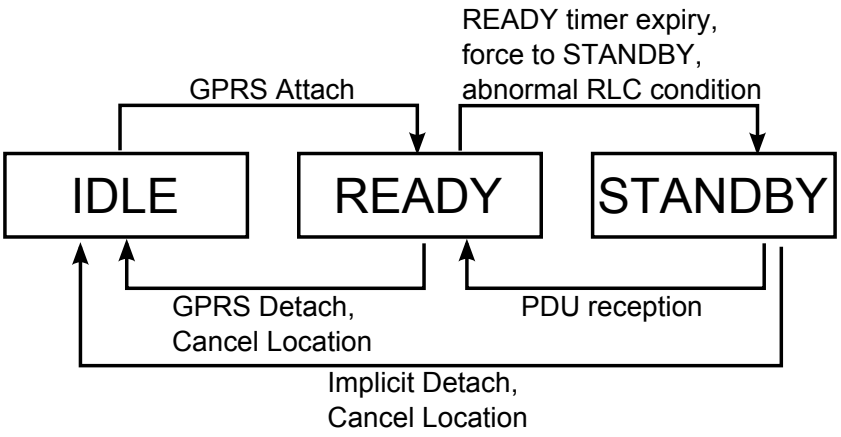


Figure 4: SGSN Mobility Management State Model.

First, consider the Mobility Management state machine depicted in Figure 4, defined in [10], and held in both the SGSN as well as the mobile device. It describes the general state of the data connection, and switches states based either on an idle timer, or when new packets arrive for the mobile device. Therefore, it also controls tunnel management, as the involved GPRS Detach and Attach procedures involve deleting and creating contexts. We identify user traffic dynamics as one vector to influence core network signaling, similar to the observations in [4].

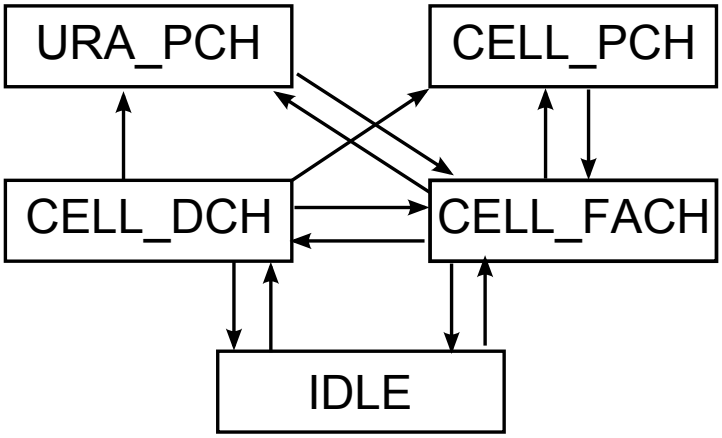


Figure 5: Radio Resource Control State Model.

The RRC state machine shown in Figure 5 governs the usage of radio channels, i.e. spectral and temporal usage of the wireless interface. State changes happen depending on the inter-arrival time of user packets. In this case, if the state machine transitions

to the IDLE state, the RAB on the path between mobile device and SGSN is not needed anymore and will be deleted, in most cases destroying the SGSN and GGSN PDP Context as well.

3.3 Discussion of GTP Signaling

Looking at the Create, Update and Delete PDP Context Request and Reply message pairs we can already deduce a certain amount of information. Measuring the time delta between corresponding Create and Delete events obviously gives the total duration a tunnel was established. Given the amount of user-plane IP traffic transferred, when the tunnel durations are short, we can expect the number of tunnel creation and deletion events to go up instead, resulting in a higher volume of signaling messages and an increase in processing for these messages. Conversely, longer tunnel durations cause an increased overall memory footprint in the involved nodes to store the PDP Contexts. Large numbers of update messages, especially combined with frequent Radio Access Technology Type (RAT) switches, are usually an indicator for highly mobile devices switching their routing area. This mobility behavior could be investigated by evaluating the update messages.

As discussed, most of the actions in the network as well as in the mobile devices are reflected in the presented tunnel management messaging. Therefore, taking a look at the dynamics of this control aspect in real networks gives valuable insights on the influence of many of the networks' aspects.

4 Network and Monitoring Setup

For our analysis, we use the Measurement and Traffic Analysis in Wireless Networks (METAWIN) monitoring system developed in a previous research project. The location of the measurement probe within the core network is highlighted in Figure 1.

As said before, the GGSN acts as the IP-layer gateway for the user traffic. It is in charge of setting up, maintaining, and tearing down a logical connection to each active MS. This logical connection, the PDP context, is conceptually similar to a dial-up connection. During set up, an IP address is dynamically assigned to the MS.

In the network under study, a so-called *direct tunnel* setup might be used for MSs connected via UMTS/High Speed Packet Access (HSPA). Such a setup consists of a direct link between GGSNs and the RNCs which is used for transporting user-plane data traffic only. Since signaling procedures such as mobility management are carried out by the SGSNs, a GGSN always has to send signaling packets via the path involving the SGSNs. Monitoring the Gn interface thus gives us access to both wide-area mobility signaling (not analyzed in this report) and signaling related to user-plane IP traffic (which we want to scrutinize). For more information on the Third Generation (3G) network structure, please refer to [12].

The METAWIN monitoring system extracts and correlates information from the lower layers of the 3GPP protocol stack, and specifically the GTP protocol on the Gn interface [13]. This includes the Radio Access Technology Type (RAT) identifier as well as the

terminal types of the mobile clients. The latter is determinable by the Type Allocation Code (TAC) part of the International Mobile Equipment Identity (IMEI) (cf. [14]) and will be discussed later in detail.

To meet privacy requirements, the METAWIN system anonymizes captured data on the fly. The packet capturing hardware deployed within the METAWIN monitoring system is synchronized using Global Positioning System (GPS). Accordingly, the packet timestamps have an accuracy of ±100 ns or better [15, p.97-98].

5 Evaluation

In this section we attempt to shed some light on the overall control plane dynamics in a mobile core network. We evaluate a dataset recorded in a live 3G network for PDP Context durations, and attempt to show the possible impact of certain device categories on the total tunnel durations. As discussed before, this can serve as a proxy metric for the signaling load on the system.

5.1 Dataset Description

Our dataset was recorded in 2011 using the METAWIN monitoring system. It consists of seven days of aggregated flow-level data for the user traffic and a summary entry for every GTP Tunnel Management transaction, the latter representing the data base for this report. It was tapped at one of the GGSNs of the operator, and contains about half of the total traffic volume handled by the operator in this period. The GTP data contain the response codes for each transactions. With these codes, failed interactions can be sorted out and treated separately.

We fed the records into a SQL database, and conducted further evaluations through scripted queries on the database. Any privacy-relevant data, e.g. the IMEI, Mobile Station Identifier (MS-ID) and any IP address involved, is only visible as hashes and can only be processed in anonymized form. Individual device types can be identified in form of the unhashed TAC on every entry. Since the hashing of the IMEI is consistent, user traffic flows and the GTP data can be cross-correlated despite anonymization, giving the opportunity for further research.

5.2 Factors Influencing Tunnel Durations

With such a dataset available and with the intent to evaluate core network signaling by looking at tunnel durations, let's first discuss some of the factors that influence this duration.

One factor are the mobile devices themselves. The device decides when it should establish a mobile data connection, how long the connection is held, or which mobile technology takes preference. Devices can be further differentiated by their operating system and their firmware (sometimes called *baseband*) which usually takes care of much of layers 1 and 2.

Some specific tunnel durations could stem from the TCP/IP stack implementations in the operating systems of the devices. TCP timeouts might be configured to different default values in different releases of OSs. Also, mobile network firewalls have been found to interfere with transport and application-layer timeout and keep-alive or heartbeat mechanisms on mobile devices [16].

Of course, the applications that run on top of the OS and generate the actual user-traffic patterns play a role as well. An example for how applications can influence network signaling is the casual game "Angry Birds" mentioned before. Since the application ecosystem for smartphones is extremely rich (and grows still), we cannot pinpoint individual ones from our aggregate dataset.

An additional factor in the picture is the user and his or her behavioral patterns. They express themselves both in the traffic dynamics and in the mobility pattern, but they are rather difficult to distinguish in such a dataset given the large amount of data and the difficulty of correctly correlating tunnel management messages. We leave this as a potential future work.

We also expect the mobile network and its protocol implementations to express themselves in the measurements. For example, the RRC idle timer is typically in the range of 10 to 30 minutes, which could mean there will be a large number of tunnels with a duration in this range. Such choices are usually made either by the mobile network operator or the device manufacturer and can vary from one implementation to another. It is therefore quite difficult to give any hard numbers in advance, and one has to correlate such aspects with certain events in the results.

Based on these factors, it was decided to make a first categorization according to the device type, be it either a smartphone, a regular or feature phone, or one of the many 3G dongles or mobile routers. Second, we also differentiate based on the device operating system, if known. Both differentiating aspects should prove valuable for example in deciding if currently some phone types put more signaling load on the network and to direct measures to improve this situation. Pitfalls in this differentiation are described in the next sections.

5.3 Difficulties of Device-based Evaluations

In our dataset, the TAC field is provided in cleartext, whereas the IMEI is only available in hashed form to preserve the privacy of device owners. The TAC is contained in the first eight decimal digits of the IMEI, uniquely identifying each device type [14]. The rest of the IMEI constitutes the serial number of the involved devices.

TACs are managed by the GSM Association which in turn assigns local organizations, distinguished by the first two digits of the TAC as Reporting Body Identifier, to allocate TACs to manufacturers. For reasons beyond us, this allocation information is not freely available. Commercial databases exist, but this is neither affordable for research institutions, nor is it conducive to our goal of providing information to the public. While there are some websites that allow one to query for specific TACs for non-commercial purposes, only very few efforts to collect TAC information into a database are publicly available. We based our data-mining efforts on a set from [17], with some additional

Table 1: Relative TAC Statistics.

	Portion of devices with entry in TAC DB
# of Flows	99.72%
Ratio of Traffic	99.97%
# of Tunnels	87.57%
# of GTP Signaling Msgs	90.95%
# of Distinct MS-IDs	80.93%

devices with known TAC collected from various sites, friends, and colleagues. Since the unit identification part of the IMEI is just six decimal digits long, popular devices will even be assigned more than one TAC, making the acquisition of all relevant TACs even more complicated.

5.4 Device Classification

For our investigation, we went through large portions of the TACs present in our dataset, and identified and categorized the most important entries. In this case, importance means various metrics like the traffic volume, the number of flows, and the number of GTP signaling messages for each TAC.

After having available the device names for most TACs, we were able to add meta-information to the entries in form of the following categories:

- The device type. We distinguished between smartphones, regular mobile phones and feature phones, and 3G USB dongles or 3G/WiFi routers.
- The operating system of the device (if known), such as Android, iOS, Series 40, BlackBerry OS etc. This is especially interesting to identify potential differences in the core network signaling patterns of devices. Note however that we cannot link USB dongles and OS types from the TAC.

5.5 TAC Statistics and Evaluation Validity

It is important to know whether our TAC mappings provide sufficient useful data to allow for the envisioned device discriminating statistics. Therefore, Table 1 provides some statistics on our knowledge of devices in the dataset. About 80 percent of all distinct devices active could be identified. Looking at the total number of GTP signaling messages, we see that we can determine the device name of over 90 percent. The flow data shows an even clearer picture, as we can identify almost all of the devices involved.

After applying the categorization to the TACs we evaluate the device composition in the network. The two largest portions of devices are smartphones and 3G dongles, while classic cell phones do not seem to play a major role anymore.

Initially, one planned endeavor was to investigate possible peculiarities of business phone behavior, especially of those easily identifiable Blackberry OS phones, but the number of distinct Blackberry devices in the dataset is too low to draw conclusions of any significance.

One observation across all device types is that about 18 percent of all mobile devices have activated their mobile data service and have signaling traffic, but do not cause any use plane traffic.

The difference between 3G dongles and smartphones is also noteworthy. While the former cause large amounts of user plane traffic (compared to the device numbers), they are responsible for but a low number of core network signaling events and tunnels. This picture is reversed for smartphones.

5.6 PDP Context Durations

Our measure of choice are the PDP Context Durations as they carry lots of meaning in being directly related to the signaling amount in the network. Therefore, we now direct our attention at the tunnel durations in the individual device and OS categories as identified via TAC values.

5.6.1 Tunnel Durations by Category

Figure 6 shows the empirical cumulative distribution functions for the PDP Context durations in our dataset. We distinguish the total duration distribution as well as the the distributions for smartphones, regular phones, and 3G dongles. It can be observed that tunnel durations range between seconds and more than one week[1].

The median differs between device types, being much longer for 3G dongles than for mobile phones. This can probably be expected, as typical dongle sessions might involve working at a laptop for periods longer than a few seconds or minutes. Also for the dongles, we observe less extremely long tunnels with durations above several hours. Again, we could hypothetically relate this to a usual laptop working environment, where the device is used for a few hours but then shut down. With this, the PDP Context is deleted as well. Interestingly, the median duration of regular phones is higher than that of smartphones. This may indicate that smartphones regularly (and perhaps automatically) cause data traffic and therefore tunnels to occur. We conjecture this to be a first indication of the "Angry Birds" effect of automatically transferring small amounts of data, e.g. weather reports, stock exchange data, RSS feeds, or email notifications. We also observe two distinct steps, one at 6.8 seconds for dongles, and one at 30 minutes in the overall and smartphone distributions. While we do not have a plausible explanation

[1] Although our dataset is one week long, some tunnels started before the beginning of that week, and ended within it. Since the tunnel start dates were still available from the system, we chose to include the data.

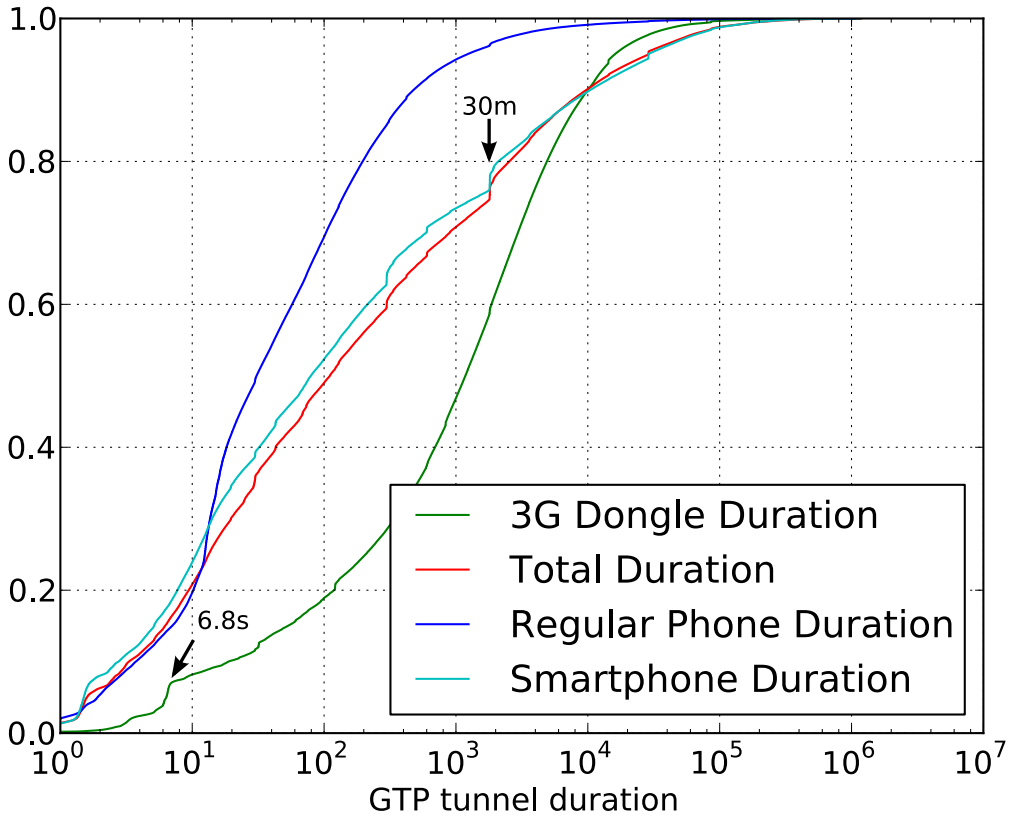


Figure 6: Tunnel duration distribution, separated for 3G dongles, smartphones and regular phones with medians at 115s (Total), 31s (Regular), 82s (Smartphone), and 1207s (3G Dongle).

for the former, the latter could be explained by a value chosen for the RRC state machine transition to the IDLE state (cf. Figure 5).

Taking an even closer look at the smartphone device fraction, we can still observe major differences as depicted in the empirical cumulative distribution functions of Figure 7. The tunnel duration distribution of the Symbian device fraction behaves much closer to the regular phones already depicted in Fig. 6. A possible explanation could be the user-base being more traditional, or the devices being feature phones whose behavior clearly differs from smartphones.

Again, a number of steps are visible in the distributions. Those steps that are only visible in one operating system type point to a source involving the phone rather the network. This especially includes the 30 seconds, 300 seconds, and 600 seconds steps (i.e. accumulations of incidents) for Android, and the 600 seconds step for iOS devices. However, whether this behavior should be attributed to the operating systems themselves cannot be decided by only looking at these distribution. Other factors, e.g. the device's firmware version and user traffic dynamics need also be observed. We leave this point

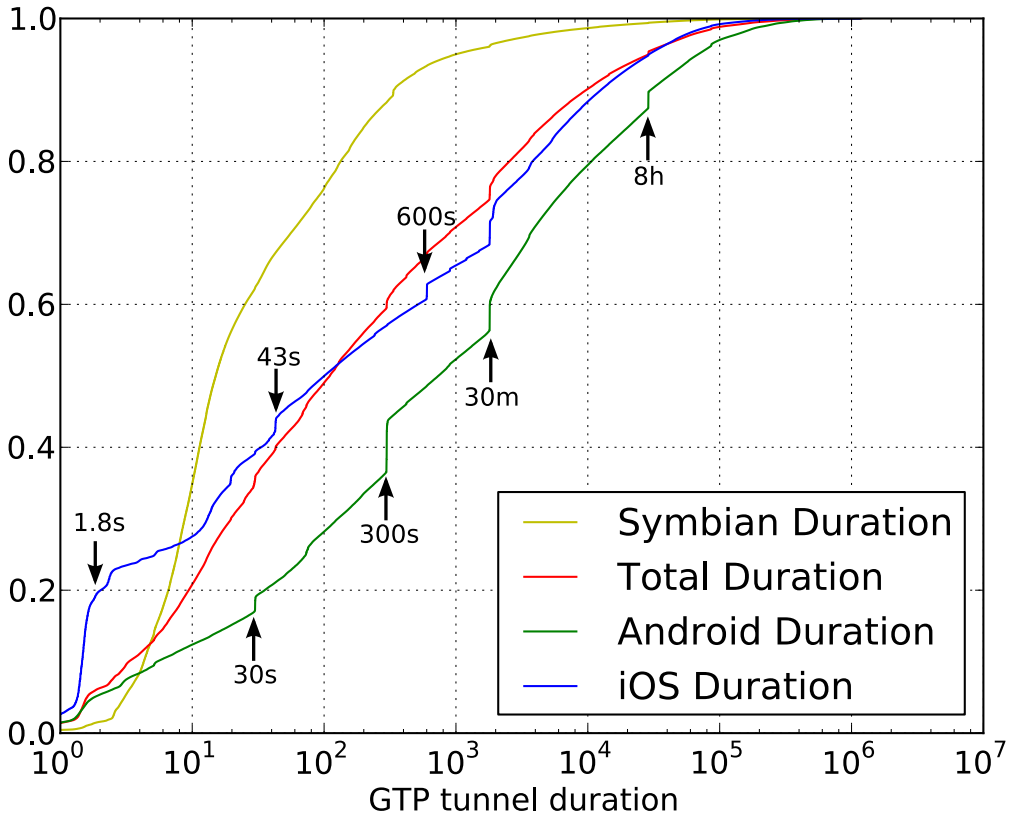


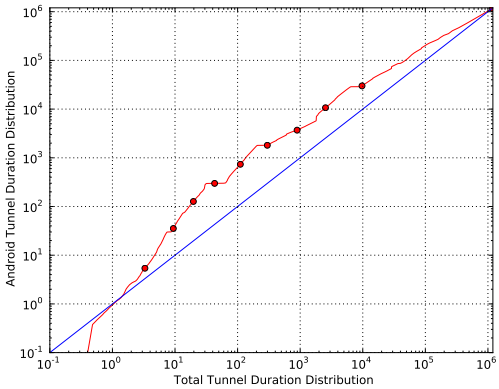
Figure 7: Tunnel duration cumulative distribution function, separated for Android and iOS devices; Medians at 115s (Total), 15.5s (Symbian), 104s (iOS), and 765s (Android).

for future work..

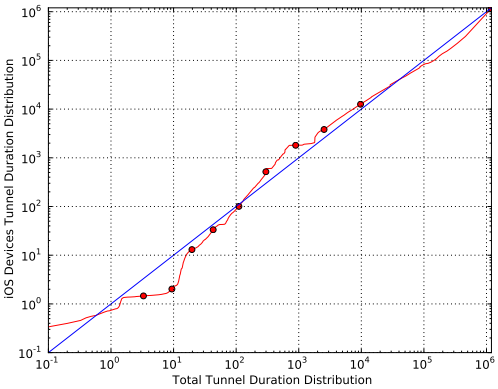
A last artifact of note are the larger number of iOS devices with very short tunnel durations. Over 20% of all tunnels established by these devices are shorter than two seconds. While the actual cause still remains unknown, it could be an interaction between short regular traffic burst and 3GPP Fast Dormancy [18] which iOS devices are known to implement. Fast Dormancy is a technique to release radio resources more quickly. It is deemed to improve device battery life, radio signaling and radio spectrum efficiency. However, due to the earlier and more frequent transition to the IDLE state, it also could cause an increase in core network tunnel management signaling, which is probably what happened in the iOS case depicted in the CDF.

5.6.2 Impact of Categories on Total Signaling

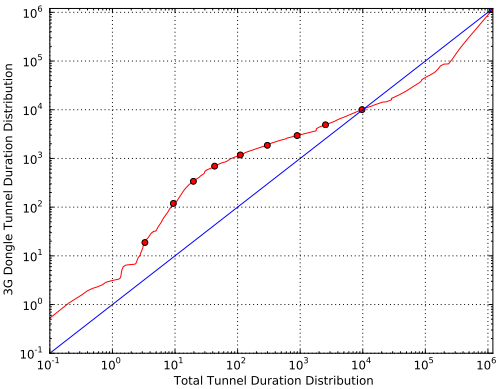
In an attempt to show which of the presented categories have an impact on the total duration (if at all), we present Q-Q plots of the various categorized durations against the total duration in Figure 8. In theory, if both durations follow the same distribution, one



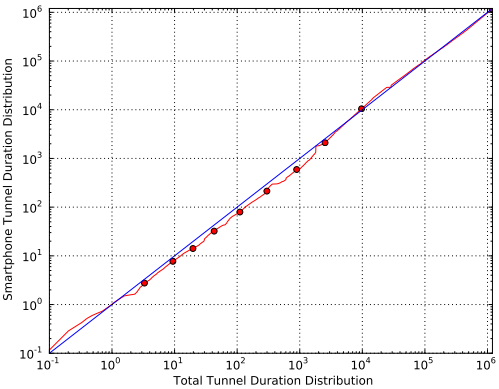
(a) Android duration distribution over the total duration distribution.



(b) iOS duration distribution over the total duration distribution.



(c) 3G Dongle duration distribution over the total duration distribution.



(d) Smartphone duration distribution over the total duration distribution.

Figure 8: Q-Q Plots of the tunnel duration distributions per operating system, with encircled deciles.

expects a straight line through the origin at an angle of 45^o. A steeper incline indicates less densely spaced values in the distribution at the y axis. Looking at igures 8a and 8b which compare different operating systems, both similar and dispersing parts can be observed. While tunnel durations on Android are more similarly distributed for the shorter and longer durations, iOS device tunnel durations are most similar to the overall tunnel duration distribution in the middle range of values.

Combining all types of smartphones together and comparing them to the other major player in any mobile network, the 3G dongles, we observe in Figure 8d that both the total and the smartphone durations are almost equally distributed (except for minor variations). On the other hand, 3G dongles follow a very different distribution, see Figure 8c. Their effect on tunnel management signaling seems to be negligible despite the large amount of traffic they are causing. Therefore, we conclude that planning and

dimensioning of the control plane needs to keep smartphone behaviors more closely in mind than that of other device types.

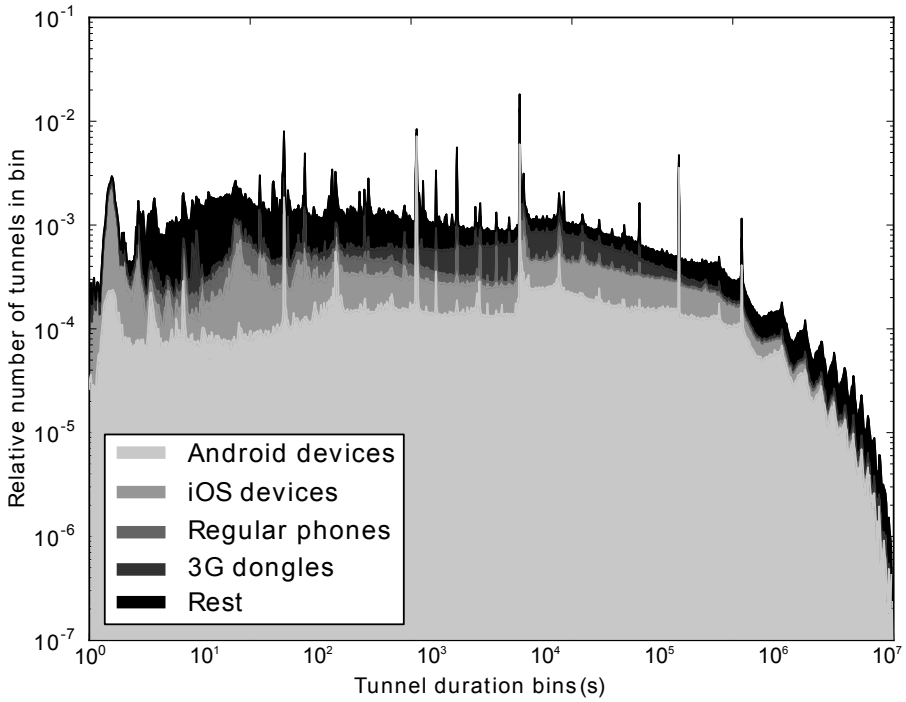


Figure 9: Stacked logscale bin plot of the number of tunnels with duration in this bin; classified by Android, iOS and 3G dongles.

Figure 9 shows another interesting influence the operating system has on signaling in the mobile core network. This plot shows the relative number of tunnels with a duration in one of 1000 logarithmically scaled bins, stacked by OS category on top of each other. As with the separate distributions, we discover that the durations are not evenly distributed, but rather follow sharp spikes. The largest spike across all categories is the one at a duration of 30 minutes, making up about 1.8% of all tunnels in the network. Since this spike happens across all device types, we think this makes a rather strong case for being network-induced, and an indication for the aforementioned possible IDLE state transition. On the other hand, the bulk in the short-to-medium ranges of tunnel duration is rather not governed by the two major smartphone operation systems but by other devices in the network, which do not show major spikes in other bins. We can also recognize a long-tail behavior in the distribution of tunnel durations.

6 Conclusion

In this report, we take a look at the signaling behavior of devices in an operational Third Generation (3G) mobile network providing Internet access. Our focus does not lie on the wireless or user-oriented parts of the network, but on signaling in the core network. To the best of our knowledge, this report is the first to offer a core network

perspective on signaling. We give a General Packet Radio System (GPRS) and Universal Mobile Telecommunications System (UMTS) network primer, and introduce GPRS Tunneling Protocol (GTP) tunnel management, explaining the causes and actions within the network.

In our observation of core network signaling involving PDP Contexts and their management based on a week long data set, we looked at the effect of device types and operating systems on the duration of GTP tunnels. We can conclude that the distribution of tunnel durations in our evaluated dataset is dominated by smartphones. This is contrary to the conventional idea that a larger volume of user plane traffic also leads to an increase of signaling. In our dataset, this would mean that 3G dongles would cause most signaling, which is definitely not the case. In this aspect, our findings support the stories of the casual game "Angry Birds" causing signaling storms in mobile networks by frequently downloading small ads, each small download resulting in disproportionate amounts of signaling load being generated. We conjecture from our results that measures taken to improve the radio interface control plane such as Fast Dormancy can have the converse effects in the core, as they could increase the tunnel churn.

All in all, our report shows that operators can determine which type of device has the most influence on the current network infrastructure by looking at and comparing tunnel duration distributions. This investigations can also lead to better network planning that is more aware of the control plane by providing the necessary tools to identify probable causes for control plane activity. Lastly, we hope to raises some awareness with programmers about the potential unintended side effects their application traffic patterns can cause.

6.1 Future Endeavors

This report serves as an introduction to the topic of the 3G core network control plane, and therefore provides only some initial insights into the actual signaling dynamics. Therefore, we would like to expand our evaluations, as there are several angles not investigated so far that could prove worthwhile.

To get a grasp of the imposed load on the network as well as the involved network nodes, a calculation of the sizes of the tunneling messages was already hinted at. To improve on this naive attempt, actual numbers on the message sizes and involved Information Elements (IEs) could be recorded in future traces. Having correct signaling traffic volume data still does not reveal the processing load on core network elements. We plan to improve our methodology in this respect by taking at a look at how long it takes for the gateway nodes to process GTP messages with respect to the current amount of user traffic and signaling. GTP tunnels also cause a certain amount of overhead through additional headers and potential fragmentation of the user traffic, providing another investigation venue for the future (albeit more oriented towards user-plane IP traffic).

Furthermore, besides the device-based classification, a differentiation based on the user traffic dynamics and correlation to signaling is planned. When looking closer at specific users, the mobility behavior also comes to mind. To investigate this, we intend to take a closer look at the occurring tunnel update messages as evidence, amongst

others for mobility.

We also look forward to searching for multiple active tunnels per device. As discussed in Section 3, the *Secondary PDP Context Activation Procedure* enables devices to establish up to ten additional tunnels attributed with a different, higher QoS level, if the network supports this. The additional load of managing and holding multiple tunnels plus the displacement of other, "lower-quality" traffic could prove to be an interesting investigation. Initial observations indicate that this feature is rarely used today by very few types of devices, but it will be of increased interest in the face of ongoing LTE/EPS deployments, whose specifications expand upon this secondary tunnel concept.

Acknowledgments

Steffen Gebert's work was done partly when the author was with the Chair of Future Communication. The stay with the Chair of Future Communication was partially supported by the research mobility support program of the FP7 project "Euro-NF" (Grant Agreement Number: 216 366).

The work was conducted during the Celtic project "MEVICO" and funded partly by the FTW strategic project "URSA Major".

The Competence Center FTW Forschungszentrum Telekommunikation Wien GmbH is funded within the program COMET - Competence Centers for Excellent Technologies by BMVIT, BMWA, and the City of Vienna. The COMET program is managed by the FFG.

References

[1] M. Donegan, "Android Signaling Storm Rises in Japan," 2012.

[2] S. Corner, "Angry Birds + Android + ads = network overload," 2011.

[3] F. Qian, Z. Wang, A. Gerber, Z. Mao, S. Sen, and O. Spatscheck, "Profiling resource usage for mobile applications: a cross-layer approach," in *Proceedings of the 9th international conference on Mobile systems, applications, and services*, pp. 321–334, ACM, 2011.

[4] P. Lee, T. Bu, and T. Woo, "On the detection of signaling DoS attacks on 3G wireless networks," in *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*, pp. 1289–1297, IEEE, 2007.

[5] F. Ricciato, A. Coluccia, and A. D'Alconzo, "A review of dos attack models for 3g cellular networks from a system-design perspective," *Computer Communications*, vol. 33, no. 5, pp. 551–558, 2010.

[6] M. Shafiq, L. Ji, A. Liu, and J. Wang, "Characterizing and Modeling Internet Traffic Dynamics of Cellular Devices," in *SIGMETRICS*, vol. 11, pp. 305–316, 2011.

[7] U. Paul, A. Subramanian, M. Buddhikot, and S. Das, "Understanding traffic dynamics in cellular data networks," in *INFOCOM, 2011 Proceedings IEEE*, pp. 882–890, IEEE, 2011.

[8] P. Svoboda, F. Ricciato, E. Hasenleithner, and R. Pilz, "Composition of GPRS, UMTS traffic: snapshots from a live network," *IPS MoMe 2006, Salzburg*, vol. 4, pp. 42–44, 2006.

[9] X. He, P. Lee, L. Pan, C. He, and J. Lui, "A Panoramic View of 3G Data/Control-Plane Traffic: Mobile Device Perspective," in *Proceedings of IFIP/TC6 Networking 2012*, Proceedings of IFIP/TC6 Networking, May 2012.

[10] 3GPP, "3GPP TS 23.060 General Packet Radio Service (GPRS); Service description; Stage 2," 2012.

[11] 3GPP, "3GPP TS 29.060 GPRS Tunnelling Protocol (GTP) across the Gn and Gp interface," 2012.

[12] J. Bannister, P. Mather, and S. Coope, *Convergence technologies for 3G networks: IP,UMTS, EGPRS and ATM*. John Wiley and Sons, May 2004.

[13] 3GPP, "3GPP TS 129.060, version 10.4.0: Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); General Packet Radio Service (GPRS); GPRS Tunnelling Protocol (GTP) across the Gn and Gp interface," January 2012.

[14] 3GPP, "3GPP TS 23.003 Technical Specification Group Core Network and Terminals; Numbering, addressing and identification," 2012.

[15] S. F. Donnelly, *High precision timing in passive measurements of data networks*. PhD thesis, Waikato University, New Zealand, June 2002. Dissertation.

[16] Z. Wang, Z. Qian, Q. Xu, Z. Mao, and M. Zhang, "An untold story of middleboxes in cellular networks," in *Proceedings of the ACM SIGCOMM 2011 conference*, SIGCOMM '11, (New York, NY, USA), pp. 374–385, ACM, 2011.

[17] C. Mulliner, "Public Research TAC Database."

[18] G. Association, "Fast Dormancy Best Practises," 2011.