Hindawi Publishing Corporation
Journal of Computer Networks and Communications
Volume 2014, Article ID 526231, 13 pages
http://dx.doi.org/10.1155/2014/526231

*Research Article*

# Exploratory Analysis of a GGSN's PDP Context Signaling Load

**Florian Metzger,[1] Albert Rafetseder,[1] Peter Romirer-Maierhofer,[2] and Kurt Tutschku[3]**

[1] *Future Communication Research Group, University of Vienna, Währinger Straße 29, 1090 Vienna, Austria*
[2] *FTW Forschungszentrum Telekommunikation Wien GmbH, Donau-City-Straße 1, 1220 Vienna, Austria*
[3] *School of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden*

Correspondence should be addressed to Florian Metzger; florian.metzger@univie.ac.at

Received 7 August 2013; Accepted 6 January 2014; Published 13 February 2014

Academic Editor: Liansheng Tan

Copyright © 2014 Florian Metzger et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper takes an exploratory look on control plane signaling in a mobile cellular core network. 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. In an investigation of core network data we take a look at statistics related to GTP tunnels and their signaling. Based on the results thereof we propose a definition of load at the GGSN and create an initial load queuing model. We find signs of user devices putting burden on the core network through their behavior.

## 1. Introduction

The Internet has reached ubiquity some time ago. Even if there is no wired access nearby you can rely on WiFi hotspots and cellular networks for wide-area coverage. These cellular networks are usually based on Third Generation Partnership Project (3GPP) specifications which have evolved from the circuit switched Global System for Mobile Communications (GSM) network into the fully packet switched Long Term Evolution (LTE) currently being rolled out. But being packet switched does not mean that it shares a lot of similarity with a typical wireline Internet protocol stack and network infrastructure. A "3G" network (a term synonymous for the typical type of cellular network used today) is very distinct from typical wired networks as it must provide, amongst other things, mobility and authentication in its core specifications rather than as optional on-top services as is typically used in the Internet.

The TCP/IP stacks largely follow two principles: "keep it simple, stupid" (KISS) and the end-to-end principle [1], which essentially means to restrict the protocols to the necessary bare-minimum and keep state only in the end systems. 3G takes a different approach, keeping a large amount of state at the obligatory nodes in its "core network," which explicitly communicate by signaling procedures defined by the 3GPP. The adverse effects of state keeping in network devices have been known a long time. For example, in the early 2000s, Internet users, running BitTorrent with connections to many peers across low-end home routers, suffered from poor performance. 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 and additional traffic on network paths and increases the number of states to be held in memory on the core network nodes. All of these attributes can be subsumed under the term "network load" which we plan to investigate in this work.

While other publications look at the near-edge interactions in these networks, research on the core is scarce, the reason for it being simple: you cannot do research without data from the operator there. Research at the edge, beginning at the IP stack level and upwards, can be conducted relatively simple. Writing simple tests and measurement scripts, often involving tcpdump and other tools, is usually all you need. But a mobile phone does not let you peek inside its layer 1 and 2 interactions (or even the implementation). Any information on this black box must be indirectly inferred from above

(forcing behavior known from the specifications through scripts) or below (spectrum analysis using software defined radio approaches). To take a look at the core's view of traffic and data, one needs access to a dedicated measurement and capturing infrastructure placed inside the network. With this, researchers can not only just look into user traffic flowing through the network but also quite easily observe the signaling-heavy mobile network control plane.

Operators usually dimension their networks in relation to the occurring user traffic. But in such a signaling-dependent architecture this might not be useful anymore, as every user traffic has to be explicitly allowed, set up, and metered through all of the network's components. This has already led to trouble in some mobile networks. User traffic tunnels, despite carrying very little actual traffic, were the cause for disproportionate amounts of signaling traffic due to being closed and reopened at a high rate. This was the unintentional cause for a DDoS in the radio access network [2, 3]. This 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 this behavior from a network that pretends to transparently carry IP.

In this paper we attempt to give some insights into the mobile network control plane and its impact on dimensioning and load modeling. To do this, some important aspects of the 3GPP specifications have to be explained to give some basic vocabulary for the following exploratory research. We then look into signaling with a focus on Packet Data Protocol (PDP) Contexts and their management through GPRS Tunneling Protocol (GTP) tunnel management procedures. Using a weeklong dataset from a mobile operator recorded at the Gn interface between the Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN), we attempt to find criteria influencing signaling. Moreover, we are formulating hypotheses on the load impact of signaling, backed by statistics gathered from the dataset.

The rest of the paper is structured as follows. Section 2 discusses relevant work in the field. Section 3 briefly introduces UMTS and GTP basics and protocol details relevant to core signaling. Section 4 gives an overview on the METAWIN data acquisition platform and a description of the dataset specifics and our approach to evaluation. While Section 5 presents a statistical evaluation of aspects of our dataset, Section 6 is an attempt on deriving an initial and simple toy load model from these statistics. Section 7 concludes the paper and gives a short outlook.

## 2. Related Work

Previous academic endeavors concerning themselves with core network signaling statistics are scarce and roughly belong to one of two areas. This includes inferring control plane behavior by either application layer active measurement at the mobile device, synthetic traces, or traces from other radio networks. Additionally, past research also evaluated actual 3G core network traces for user traffic characteristics. This work is also an extension to our research report [4] aiming to provide more in-depth statistical analyses to the control plane.

Stories about signaling storms and overloaded control planes in mobile networks [2, 3] blame specific combinations of device types, operating systems, and applications to cause excessive amounts of signaling in the radio network. Many popular free-to-play mobile games use periodically display refreshed advertisements. This leads to a scenario, wherein a large amount of devices constantly set up and tear down data connections just to retrieve new ads, thereby triggering tens of Radio Resource Control- (RRC-) related control plane messages on each retrieval and straining the signaling-heavy structure of current mobile networks. These dynamics are already under investigation by several publications. A paper on cross-layer interaction in mobile cellular networks falls into this category [5], discussing interaction, for example, between application layer and RRC 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 [6], mobile network traces are used to simulate a malicious signaling storm by transmitting low-volume user-plane traffic with interdeparture 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 a possible scale of this type of denial-of-service attack. For investigating the transition of RRC states, [7] proposes simple yet effective application layer based methods. This is further enhanced by research from Schwartz et al. [8] using this technique to analyze the radio signaling load and thus power efficiency from different applications.

While the approaches above concern themselves with radio signaling they neglect core network signaling. The following research papers have access to core network measurements but do not directly tackle signaling. The authors of [9, 10] both take the approach of looking at high-level user traffic characteristics in a mobile network, focusing on temporal and spatial variations of user traffic volume and peeking at the influence of different devices on this metric. Additional user flow and session traffic metrics are being looked at in [11] concluding that flows are generally shorter than in wired networks, with a potential impact on signaling load. Svoboda et al. [12] conducted a core network measurement study of various user traffic related patterns and also provided an initial insight into PDP Context activity and durations. A recent publication provides an investigation aimed at RRC signaling between Radio Network Controller (RNC) and SGSN [13]. The authors classify their evaluations based on device model and vendor and on the application type and find that different devices strongly differ in their RRC characteristics, which could possibly also have an impact on GTP signaling. A 2010 publication [14] indirectly infers RRC signaling and deducts that the involved RRC state machine is largely inefficient in terms of signaling overhead and energy consumption for typical traffic patterns seen in the data. The authors of [15] give us some thoughts on the influence of core network elements on one-way delays in mobile networks and the expected load impact of these elements. A final paper discusses some theoretical

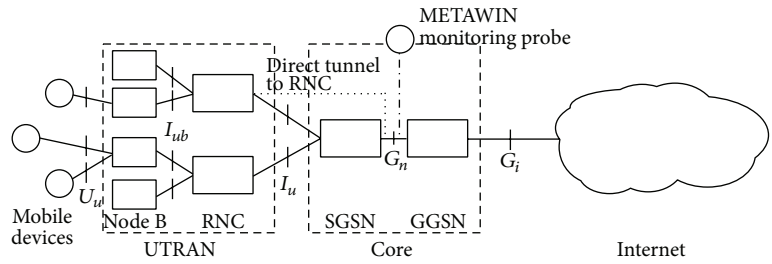



Figure 1: Simplified setup of the packet switched domain in an UMTS network including a METAWIN monitoring probe.

denial-of-service (DoS) attack scenarios on mobile networks [16]. A DoS typically needs to find a (performance-wise) weak link in an architecture or employs an amplification attack. This presents helpful information in evaluating core network load and finding bottlenecks.

All of these publications touch parts of the areas tackled in this paper but do not yet present a complete picture. We think that the combination of the focus on core signaling, a statistical evaluation of PDP Contexts with an investigation of sources influencing these, and a simple load model are genuine contributions of our work.

## 3. GPRS and Tunnel Management

This section serves as a short introduction on cellular data network basics and describes relevant details of GPRS Tunneling Protocol (GTP), the tunneling protocol under investigation.

*3.1. GPRS Fundamentals.* The packet switched domain of a Universal Mobile Telecommunications System (UMTS) network is an evolution of General Packet Radio System (GPRS) and thus closely related to it. First defined by the Third Generation Partnership Project (3GPP) in Release 99, it 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 [17] defines the basic aspects involving GPRS protocols and its system architecture. TS 29.060 [18] describes the specifics of GTP flowing across the Gn and Gp interfaces which forms the foundation for our work.

As shown in Figure 1, user traffic originating at any Mobile Station (MS) connected to the radio network flows through a Node B (also called base station), which provides radio connectivity. The traffic of multiple Node Bs in the same area is aggregated into a Radio Network Controller (RNC). These base stations 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 to serve as mobility anchor for mobile devices. It is also the endpoint for Radio Resource Control- (RRC-) based signaling and the Radio Access Bearer (RAB), the radio counterpart to the core network user traffic tunnel. The GGSN provides the gateway to the public Internet. The Gn interface connects those two nodes, using the GTP protocol to encapsulate 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 in the core.

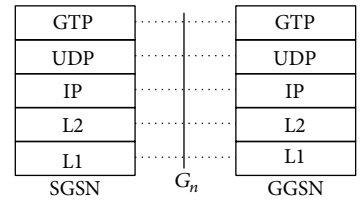



Figure 2: Typical signaling protocol stack at the Gn interface between SGSN and GGSN.

*3.2. GTP Signaling.* Tunnels state is held in the SGSN and GGSN as Packet Data Protocol (PDP) Context data structures. These contain various information, such as the device IP address, International Mobile Subscriber Identity (IMSI), and a tunnel identifier. The concept is used to isolate user traffic from core network control plane signaling and to provide certain Quality of Service (QoS) guarantees to the user traffic. Multiple QoS profiles per device can also be established by setting up up to ten secondary contexts beyond the primary PDP Context. However, QoS secondary contexts are very rarely in use today; any user-plane IP traffic is typically transported within the primary "best effort" tunnel.

The GTP-C signaling, responsible for the context management interactions, contains procedures for managing data paths, MS locations, mobility, and, of course, tunnels. GTP messages usually come as request-response pairs. Neither part has fixed size but is rather constructed from a number of Information Elements (IEs), many of which are either optional or variable length through additional optional fields.

The focus of our work will be the three tunnel management message pairs involved in the maintenance of PDP Contexts. These are as follows.

(i) *The Create Context Message*. It is part of several larger control plane procedures that activate the GTP tunnel for a mobile device. These can be initiated from the network as well as the device itself, again depending

on the specific implementation of the architecture. 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 failures include failed user authentication, lack of resource, or unrecoverable system failures.

(ii) *Delete Context Message*. This indicates the immediate release of the Context involved. Together with the Create event, these mark the beginning and the end of every GTP tunnel, making them good candidates to determine tunnel durations for our load evaluations.

(iii) *Update Context Messages*. Several procedures also emit tunnel update messages, when some aspect of the tunnel has changed, for example, occurring in mobility and load-balancing related procedures but also procedures involving secondary tunnels for a device. 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.

The variable-length nature of these messages makes evaluating the imposed network signaling overhead 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 lower bound for the message size with all IE present is 307 bytes.

Taking the maximum size we arrive at a naive estimate of the maximum overhead on user traffic induced by tunnel management signaling in our dataset. The estimated ratio of (tunnel management) signaling traffic to total user-plane traffic in our dataset is a minute 0.10%. Therefore, the volume of control plane traffic appears to be noncritical in this setup. Thus, we assume that the overload problems mentioned in related work arise rather in other areas affected by control plane signaling. This includes the memory profile of the control plane state kept in the gateway nodes, the time required to process the large quantity of information held in the messages, or the imposed latency through several message round trips during transactions.

*3.2.1. State Machine Influences on GTP.* As indicated, most nodes in a cellular mobile network keep all sorts of state characterizing the data connection. For the tunnel management aspects, two state machines are of special note, namely, the mobility management and RRC state machines. The former, defined in [17], describes the general state of the data connection and switches states based either on an idle timer or arrival of new packets for the mobile device. The RRC state machine depicted in Figure 3 governs the usage of radio channels. State changes happen again depending on user activity and inactivity. Based on the state both procedures can enable and disable radio tunnels as well as core network tunnels, making them a good example of user traffic dynamics directly influencing core network signaling, similar to the observations in [6].

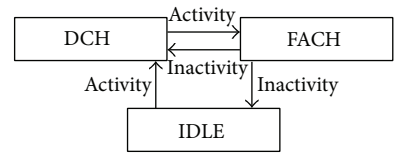



Figure 3: Simplified Radio Resource Control state model.

*3.3. Discussion of GTP Signaling.* 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.

Looking at the Create, Update, and Delete PDP Context Request and Reply message pairs we can already directly deduce some possibly load-related information. The time between a request and its corresponding response could also be an indicator for the amount of processing involved for this message as well as the current general processing load at the GGSN. The total tunnel duration originates from the time delta between corresponding Create and Delete events. A decrease in the average tunnel duration will increase the number of total tunnels and thus also the volume of signaling messages and the necessary 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 (RAT) switches, are usually an indicator for devices often switching their routing area.

## 4. Dataset And Methodology

For our analysis, we use data acquired by the Measurement and Traffic Analysis in Wireless Networks (METAWIN) monitoring system developed in a previous research project [19].

The location of the measurement probe at the Gn interface within the core network, marked in Figure 1, gives access to both wide-area mobility signaling (not analyzed in this paper) and signaling related to user-plane IP traffic (which we want to scrutinize). The METAWIN monitoring system extracts and correlates information from the lower layers of the 3GPP protocol stack, specifically the GTP protocol on the Gn interface [20]. This includes the Radio Access Technology (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. [21]) and will be discussed later in detail.

To meet privacy requirements, the METAWIN system anonymizes captured data on the fly at multiple layers: the application-level payload is removed and all user identifiers (e.g., IMSI) are hashed before recording. That is, single MSs in

our dataset may be differentiated by means of an anonymized Mobile Station Identifier (MS-ID) but not traced back to the actual customer. 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 [22, pages 97-98]. We further accommodate to the sensitive nature of this dataset by disclosing as little information as possible but as much as is required for this research.

*4.1. Dataset Description.* The METAWIN-recorded dataset used in our evaluation is a weeklong trace from the third week of April 2011. It consists of 2.2 billion aggregated flows for the user traffic and 410 million GTP tunnel management transactions, the latter representing the data base for this paper. 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 contains the response codes for each transaction. 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, for example, the IMEI, MS-ID, and any IP address involved, is only visible as hashes and is processed in a privacy-preserving manner. Since the hashing of the IMEI is consistent throughout the dataset, user traffic flows and the GTP data can be cross-correlated despite anonymization, giving the opportunity for further research.

*4.2. Device Identification and Classification.* The type of a device can still be identified in form of the TAC on every entry. The TAC is part of the IMEI, uniquely identifying each device type [21]. The rest of the IMEI constitutes the serial number of the involved devices, which is not present in the data.

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 TAC to manufacturers. However, 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 noncommercial purposes, only very few efforts attempt to collect TAC information into a publicly available database. We based our data-mining efforts on a set from [23], with some additional devices collected on our own. 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 TAC even more complicated.

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 metainformation and categorize the entries based on their device type and operating system. For the device type we partitioned the devices roughly into smartphones, regular mobile phones, and 3G USB dongles or 3G/WiFi routers. The operating system includes most of the popular incarnations found in the network at measurement time, including Android, iOS, and Symbian. Note, however, that many devices, especially USB dongles, cannot be linked to any specific OS.

As we are working with an incomplete TAC database it is important to know whether our TAC mappings provide sufficiently 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 and active devices 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 in the packet switched domain anymore. One observation across all device types is that about 14 percent of all mobile devices have activated their mobile data service and have signaling traffic but do not cause any user-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 few core network signaling events and tunnels. This picture is reversed for smartphones with much signaling and little amounts of data.

## 5. Core Network Load Statistics

Having characterized the dataset available to us we now shed some light on the control plane and load dynamics in a mobile core network and attempt to show the possible impact of certain devices or other properties of the network.

*5.1. Defining Core Network Load.* The primary question driving this investigation is, "how can load in a core network be defined and measured?" A summary of our thoughts to this question follows here.

With the basics of the architecture in mind, a top candidate for high load is the GGSN. All traffic leaving or entering the packet switched domain must go through this element, and it is in control of the described GTP signaling procedures as well. Being an endpoint for the GTP tunnel makes it responsible to sort and encapsulate incoming traffic into the corresponding user tunnel. To accomplish this a lot of state has to be kept—and processed when signaling occurs. Therefore, our working hypothesis is that, in order to determine load, the GGSN needs to be monitored closely, and any traffic related to this node should be investigated for indications of the current load.

Table 1: Relative TAC statistics.

| | Percentage of devices with entry in TAC DB |
|---|---|
| % of flows | 99.72% |
| % of traffic | 99.97% |
| % of tunnels | 87.57% |
| % of GTP signaling messages | 90.95% |
| % of distinct MS-IDs | 80.93% |

For our definition of the term "load" we differentiate between signaling traffic and overhead, on the one hand, and processing load and memory consumption on the other hand. Both are measures of load at specific nodes. While the former mostly has an impact on the actual network traffic, the latter can only be grasped inside the network element. With our data we can directly investigate the signaling traffic but indirect measures for the processing load and memory usage have to be found. In the rest of this section we evaluate the results of several approaches to both of these definitions of load.

While looking at the GGSN may be the most obvious choice, it is by far not the only one. In addition to GTP tunnels the SGSN has to handle RAB and mobility management as well. However, it is assumed that there are more regionally distributed SGSN nodes present in a typical mobile network. This means that a single element would have to handle less mobile devices and therefore load. One has also to bear in mind that the SGSN can be completely circumvented by setting up a direct tunnel between GGSN and RNC.

Apart from the two gateways directly inside the traffic path, there are several other nodes essential to the control plane decision making, which may very well be also very load sensitive. The Home Location Register (HLR), for example, is a central database storing all user related information which needs to be retrieved any time a user needs to undergo initial authentication and authorization. Typically, the procedures the elements are involved in are fewer and they are also harder to investigate with the data available to us. Hence, it was decided to concentrate just on the case of the GGSN.

*5.2. Load Influencing Factors.* Having described our understanding of core network load we can now move on to discuss some of the factors that could influence the load, making them targets for our evaluation.

The first and arguably one of the most important factors is the mobile devices themselves. Specifically, this covers the behavior of the network layer 1 and 2 implementation (sometimes called "'baseband") as well as the operating system (OS) and the running applications. The OS and baseband decide when the device should establish a mobile data connection, how long the connection is held, or which mobile technology takes preference. Depending on the access technology, be it GPRS, EDGE, UMTS, HSPA, or HSPA+, we can expect subtle differences through their specifications, for example, in the timing of the radio transmission intervals, which could influence our investigation.

Some specific tunnel duration properties could stem from the OS's IP and transport protocol implementation. For example, TCP timeouts might be configured to different default values causing mobile connections and tunnels to be held either shorter or longer. Also, mobile network firewalls have been found to interfere with transport and application layer timeout and keep-alive or heartbeat mechanisms on mobile devices [24].

The actual user traffic patterns are generated by the applications running on top of the OS. For example, the aforementioned ad-based free applications with their ad-retrieval strategy cause network traffic and possibly signaling in certain intervals. Since the application ecosystem for smartphones is extremely rich and ever growing we cannot pinpoint individual ones from our aggregate dataset.

An additional factor in the picture is the user and her or his behavioral patterns. They present 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 potential future work.

Easier to observe are the temporal effects of user behavior, which do not target individual users but the overall effects of device usage based on the time of day, the day of the week, or other time spans. In network user traffic analyses diurnal effects are typically very distinct with peak traffic some time during the day and the lowest traffic shortly after midnight. But these investigations are for user traffic only. We aim to find out if the mobile network control plane shows similar patterns and can thusly be correlated to user traffic.

We also expect the mobile network and its protocol implementations to be visible 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.

*5.3. Individual Examinations.* To examine some of these factors, we present the following number of individual investigations. Our measure of choice is the GTP tunnels as they carry lots of meaning in being directly related to the amount of signaling in the network. We investigate their duration as well as the number of arrivals and look at a measure for the processing time of events at the GGSN. These insights will also allow us to build a simple toy model for the core network load in the next section.

*5.3.1. GTP Tunnel Duration.* In our evaluation, we define the duration of a GTP tunnel as the time between a GTP Create and the corresponding GTP Delete event. After the reply for a Create has been sent from the GGSN any setup procedures at the node should have completed and it should recognize incoming traffic from or to this user. After the Delete, the

user's traffic will not be routed anymore. Any lazy cleanup happening after the Delete is not relevant for this specific investigation.

We differentiate all the tunnel events in our dataset based on two factors. First, we look at tunnels from different device types, be it a smartphone, a regular or feature phone, or one of the many 3G dongles or mobile routers. After that, we investigate possible influences from the operating system. Both categorizations 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.

*Influence of the Device Type.* Figure 4 shows the empirical cumulative distribution functions for the PDP Context durations in our dataset. We distinguish the total duration distribution as well as the distributions for smartphones, regular phones, and 3G dongles. It can be observed that tunnel durations range between mere seconds and more than one week. (Although our dataset is just 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.)

The median is clearly different 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. Again, this could be attributed to a hypothetical laptop working environment, where the device is used for a few hours but then shut down and the PDP Context getting deleted.

Interestingly, the median duration of smartphones is slightly lower than that of the total distribution. This may indicate that smartphones more frequently (and perhaps automatically) cause data traffic and therefore tunnels to occur in short and more interspersed bursts. We conjecture this to be a first indication of applications automatically transferring small amounts of data, for example, 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.

*Influence of the OS.* Taking an even closer look at the smartphone device fraction and differentiating the operating system to Symbian, Android, and iOS, we can observe additional differences as depicted in the empirical cumulative distribution functions of Figure 5. The tunnel duration distribution of the Symbian device fraction behaves much closer to the regular phones already depicted in Figure 4. A possible explanation could be the user-base being more traditional or the devices being feature phones whose capabilities clearly differ from smartphones.

Again, a number of steps (i.e., accumulations of incidents) are visible in the distributions. Those that are only visible in one operating system type point to a source involving the device rather the network. This especially includes the 30 seconds, 300 seconds, and 1800 seconds steps 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 just looking at these distributions. Other influence sources, for example, the device's firmware version and user traffic dynamics, need also to be observed.

A last artifact of note is the large number of iOS devices with very short tunnel durations. Over 20% of all tunnels established by these devices are shorter than two seconds. Our working hypothesis is that this is an interaction between short regular traffic burst and a form of Fast Dormancy [25]—a technique to explicitly release radio resources—which iOS devices are known to implement. It is deemed to improve

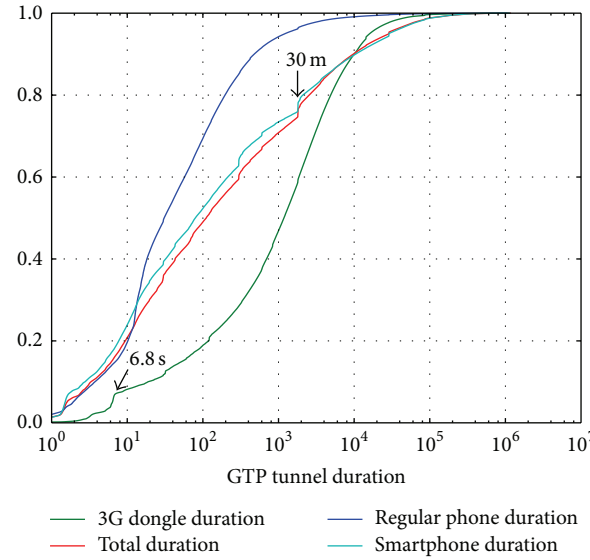



Figure 4: Tunnel duration distribution, separated for 3G dongles, smartphones, and regular phones.

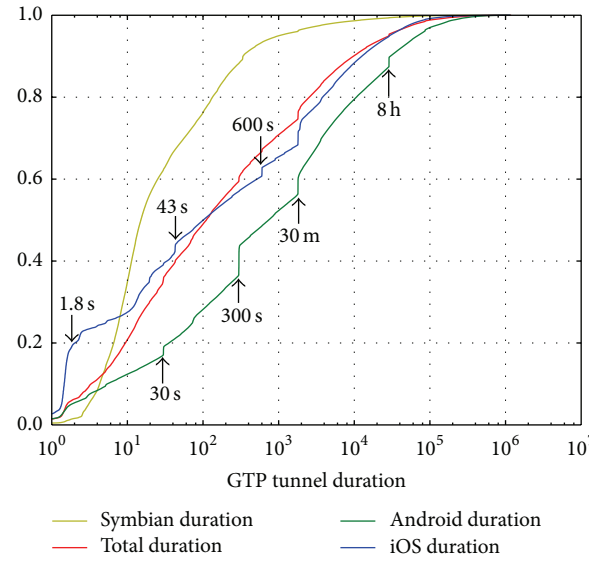



Figure 5: Tunnel duration cumulative distribution function, separated for Android and iOS devices.

device battery life, radio signaling, and radio spectrum efficiency. However, due to the earlier and more frequent radio state changes, 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.

*5.3.2. Number of Tunnel Arrivals and Interarrival Time.* While tunnel durations and the involved signaling at the beginning and end of the duration are one aspect of control plane load, the number of tunnel arrivals might be another, which we are looking into in this section.

In addition to describing the arrival process based on the number of arrivals, we also take a look at the tunnel interarrival time. Specifically, with this process we mean the arrival of tunnel requests, that is, GTP CREATE requests, at the GGSN. This also adds to the foundation of the load model constructed in the next section.

Figure 6 depicts the number of arrivals per second during the whole weeklong period. Of note is the clear bimodal nature with one peak around twelve and the other in the low thirties. While the distribution is rather compact around these two peaks, there are some clear outliers up to 107. If we again hypothesize that the increased number of arrivals means higher load in the network, we can assume that load is not constant but rather switches between two modes with some periods with extraordinary load induced by an increased number of arrivals.

To find the cause of these two modes we take a peek at the diurnal arrival pattern. Figure 7 contains a violin plot showing again the arrivals per second but broken down by time of day. A violin plot, being conceptionally similar to a box plot, additionally shows the density of the individual items on the vertical axis. The nocturnal median from around midnight to 5 a.m. and the daytime median, 8 a.m. to 7 p.m., closely resemble the two modes found in the histogram. In between are short transition phases. Notably, during daytime the arrivals and their densities are spread out on a much larger value range. This could be an indication of load fluctuations in the system, as more active users could lead to an increase in load variance.

To investigate the arrivals from yet another angle we take a look at interarrival time of the tunnels in Figure 8. This metric is more suited to describe the arrival process in the toy queuing model we propose. The empirical CDFs are again broken down by time of day, and the same diurnal load oscillation can be observed. The medians range between about 20 and 60 milliseconds. Figure 8(a) represents all tunnel requests that the GGSN handled. It shows wave-like steps in 20 ms intervals in the plot. Because this is happening in regular intervals at every time of the day, we believe that this effect must originate from a source inside the mobile network and is not induced from the outside, for example, through mobile devices.

This becomes even more peculiar when we further evaluate the tunnel arrivals. We now distinguish between active tunnels—that is, tunnels that actually transported user traffic during their lifetime (cf. Figure 8(b))—and even more specifically, active tunnels created during GPRS (Figure 8(c))

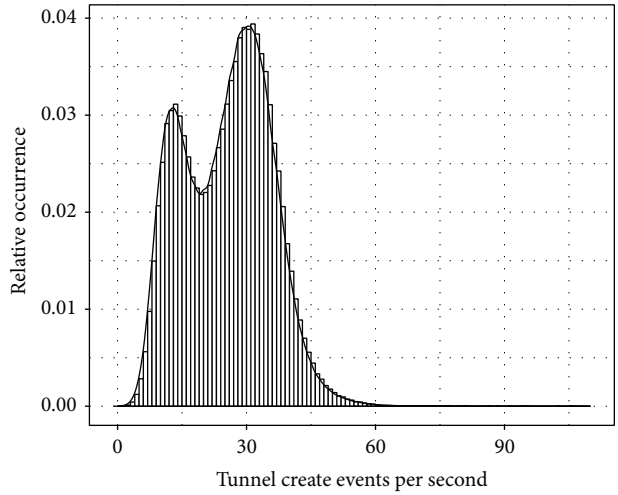



Figure 6: Tunnel arrivals in one second intervals.

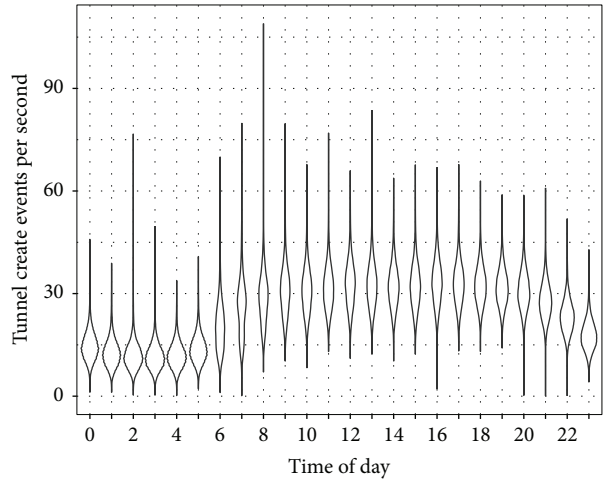
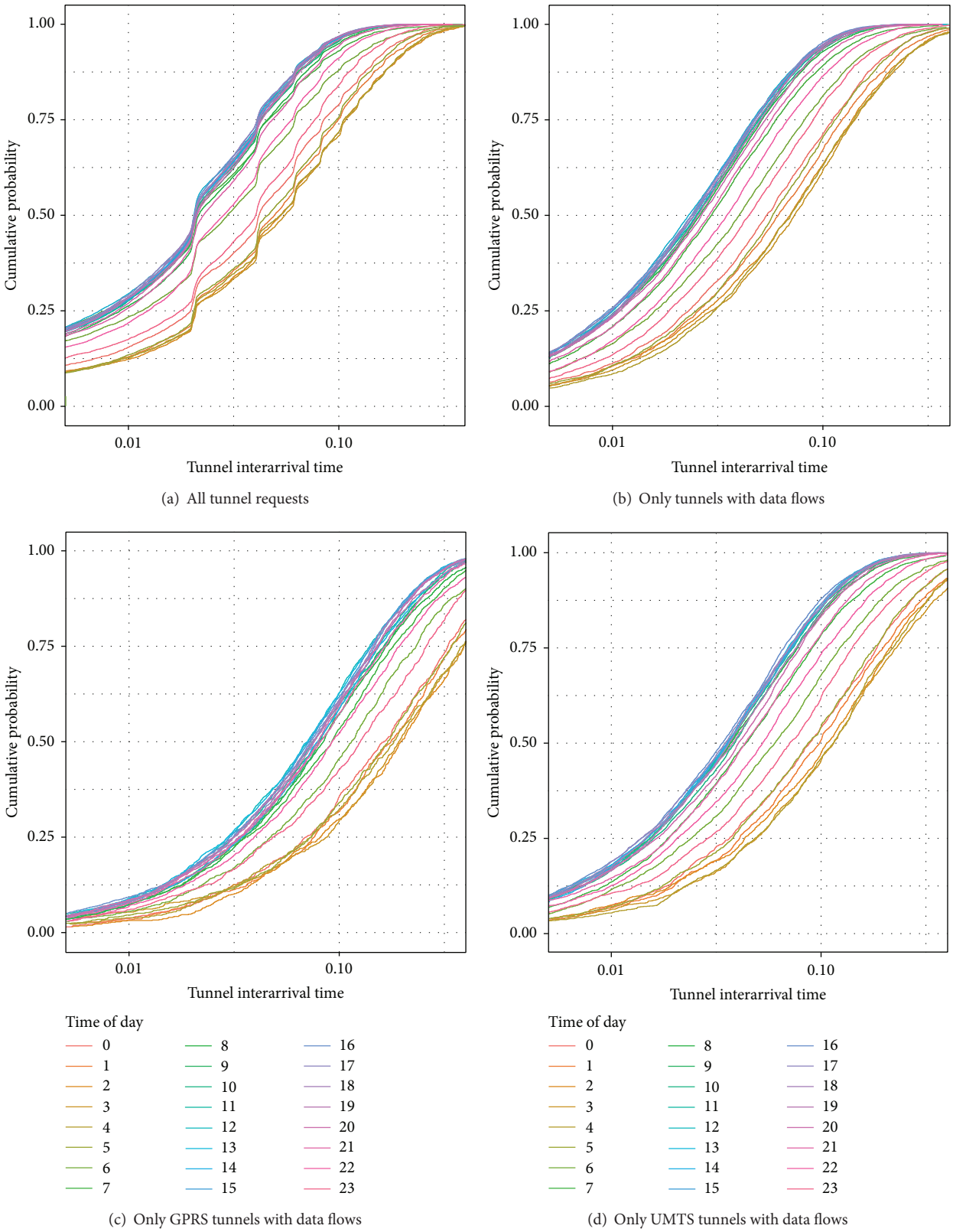



Figure 7: Violin plot of tunnel arrivals in one second per time of day.

or UMTS (Figure 8(d)) connectivity. Note that only about 86% of requested and created tunnels were actually used for user data transmissions in their lifetime. The 20 ms steps occur strongest when observing all tunnel arrivals, and in a weaker form it is also present in the active and UMTS tunnel portion.

Our working hypothesis as to the origin of the effect is the Transmission Time Interval (TTI). It determines the duration of a radio transmission and is usually either 10 or 20 milliseconds in length. It is also in sync for the whole network of base stations making the TTI noticeable even when not measuring directly at the radio link. The observed step-width of 20 ms therefore indicates that the signaling procedure the GTP Create is part of includes at least one trip from the mobile device over the radio interface. This makes sense, as the tunnel is typically created during the GPRS Attach procedure, which is indeed initiated at the user's device. Unfortunately, this also makes the tunnels arrive in batches, which could momentarily increase the load at the




(a) All tunnel requests

(b) Only tunnels with data flows

(c) Only GPRS tunnels with data flows

(d) Only UMTS tunnels with data flows

Figure 8: Empirical cumulative distribution function of the tunnel interarrival time in seconds by time of day.

GGSN that then would need to process more requests at once than if the arrivals followed a smooth stochastic distribution.

*5.3.3. Tunnel Event Processing Time.* This brings us to another and potentially more direct measure of GGSN load, namely, the processing time of update events, meaning the time it takes for the GGSN to fulfill a GTP request (Unfortunately, issues with the dataset did not allow the investigation of the processing time of create or delete messages.). This is calculated from the requested and finished timestamps of every GTP update event in our dataset. As the measurement is conducted at the Gn interface these timestamps represent the time the GTP signaling request moves to the GGSN and the time the response transitions through the link.

As stated in the previous section, it would be of special interest to know if the setup time of tunnels is influenced by anything, as this is one of the GGSN's most time-sensitive jobs and can impact the time a user has to wait before being able to actually transfer data.

However, we could investigate the processing time of GTP update messages. The core network transmits roughly two orders of magnitude more update than either Create or Delete events and therefore the number of usable events exceeded the significance level. While no direct investigation of the setup and deletion procedures was possible with these events, a rough overall picture of load can still be attained through this. Figure 9 depicts a band of empirical cumulative distribution functions for the processing time of Update events broken down by time of day. The processing time is almost uniformly distributed between 2 and 22 milliseconds, with a slightly longer duration during the evening, making for a continuous uniform distribution. This is rather unexpected as uniform distributions do not usually occur in computing processes. According to the central limit theorem one would rather expect to see a normal distribution influenced by, for example, cascaded scheduling or queuing artifacts. In the future we hope to investigate these features more closely, including a proper investigation of the tunnel setup and teardown processing time, if the dataset allows it.

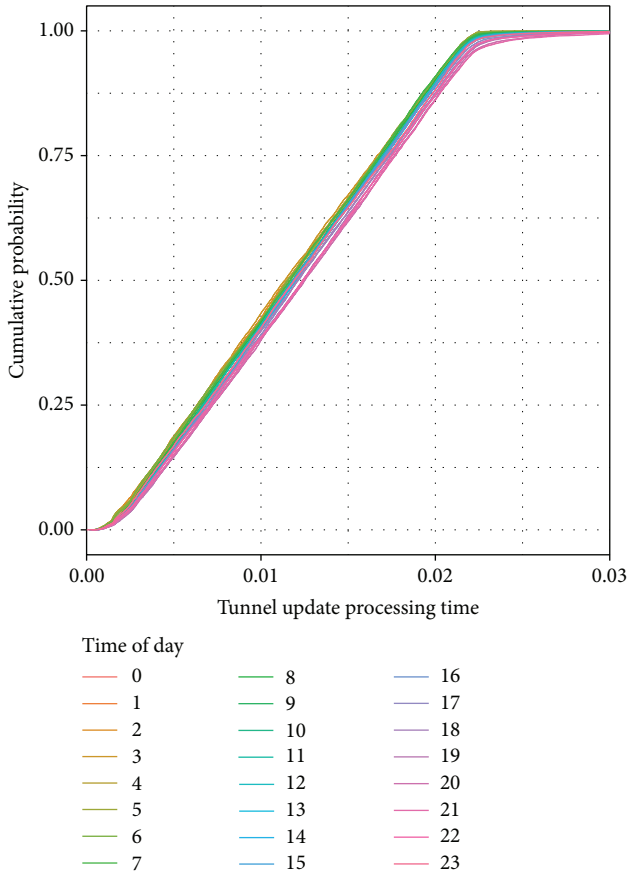



Figure 9: Empirical CDFs of the time it takes a GGSN to process a GTP update event, plotted for each hour of the day.

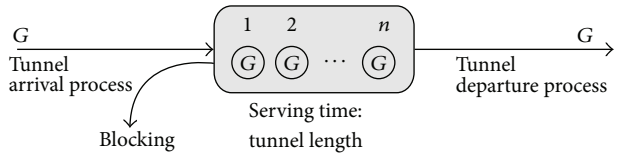



Figure 10: Simple toy model for tunnel-induced load on the core network.

## 6. Load Modeling

Drawing conclusions from statistical analysis alone is a difficult task. The next logical step lies therefore in the creation of models abstracting this real system, making them easier to calculate with the loss of some precision. This and future improved models should support network operators in predicting the signaling load in their core network with the benefit of improved network engineering and correctly scaling core components.

*6.1. Creating a Simple Toy Queuing Model.* To begin the modeling process we attempt to represent the tunnel management as a queuing system, specifically as a $G/G/n - 0$ system in Kendall's notation. Figure 10 shows this model for the case of our proposed tunnel load metric. Here, tunnels enter the system by a general random distribution, are then "served" at the GGSN for the duration of their existence, which also follows a general distribution, and leave the system, that is, are torn down, afterwards. If the serving units are filled, blocking occurs and arriving tunnel requests are rejected.

In this case "servers" correspond to available resources at one or more GGSN, making the maximum number of tunnels hard to guess and depend on a number of factors. This could include soft-limits like the specific configuration, and hard-limits, for example, the GGSN's processing and memory constraints. Unfortunately, all of these are unknown to us.

For the purpose of creating a toy model we simplify the $G/G/n - 0$ to a $M/M/\infty$ system. As stated, no actual limit to the number of concurrent tunnels is known and the data also does not show any obvious limits. Thus, we can safely assume an unlimited system and do not have to handle blocking or queuing explicitly.

Furthermore, we fitted univariate distributions to the experimental data for the tunnel interarrivals and durations and tested the goodness of the fit both numerically, using Pearson's $\chi^2$ test, and visually for the density and CDF plots. No standard random distribution reaches the significance level for either process. We attribute this fact largely to the various artifacts in the data, for example, the described wave effect every 20 milliseconds in the interarrival time. Matching them visually (confer also the empirical CDF plot in Figure 11) we find that the exponential fit is reasonably close to the experimental data in both the arrival and duration cases. Again, these distribution fits are just for a toy model to lay the groundwork for future and improved modeling.

Now, assuming both a Poisson arrival and an exponential serving process, a Markov chain representing the queue can be set up (cf. Figure 12) and stationary analysis can be conducted. From the measured data an arrival rate of $\lambda = 25.64123$ and the parameter $\mu = 0.0001586728$ for the service time distribution are calculated. Using Little's law this gives an estimate for the mean number of concurrent tunnels at the GGSN of

$$L = \frac{\lambda}{\mu} \approx 161\,599. \quad (1)$$

As stated, the amount of state held at the node and propagated through the network is directly related to the number of tunnels. Therefore, we propose this metric as an initial estimate of the load at the GGSN.

*6.2. Modeling Outlook.* On the basis of this toy model better fitting models can now be constructed. Those should also factor in more of the core network's properties and specified parameters omitted in this model. Specifically, this means shifting from $M/M/\infty$ to the more generalized $G/G/n$ and therefore finding better distribution fits for the involved processes.

It is also entirely possible that the single queue approach is not the best way to describe control plane load. Several load influencing factors discussed earlier have direct influence on the tunnel arrivals and duration, for example, the device type or the Radio Access Technology. Therefore, amongst other approaches, multidimensional queuing networks or fluid flow could be a better fit. Our plan is to conduct further investigations into the modeling of mobile core network signaling. This also includes a rough simulative approach, which could also be used to validate our models against experimental data.

## 7. Conclusion

In this paper, we took 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 paper is the first to offer an in-depth core network perspective on signaling. We gave a General Packet Radio System (GPRS) and Universal Mobile Telecommunications System (UMTS) network primer, introduced GPRS Tunneling Protocol (GTP) tunnel management, and evaluated a weeklong dataset recorded in a mobile operator's core network.

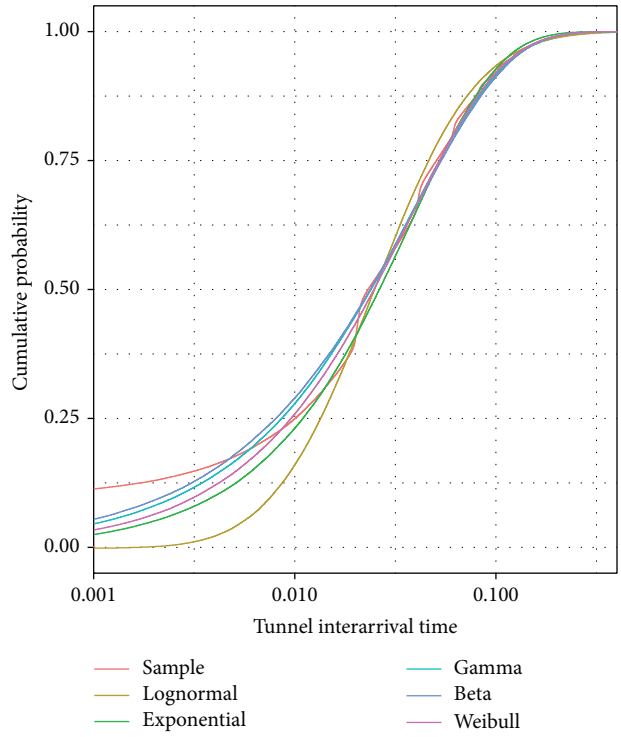



Figure 11: Empirical CDFs of the sampled interarrival time and fitted theoretical distributions.

In our observation of core network signaling involving Packet Data Protocol (PDP) Contexts and their management, 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. Conventionally, one would assume that there is a direct correlation between user-plane traffic and signaling. Our investigation shows and gives initial indications that this is not the case. Rather, our paper shows that network operators can determine load-inducing factors, for example, mobile device types, much better by looking at and comparing tunnel duration distributions and tunnel arrival signaling characteristics.

For additional load investigations we also look at the interarrival and processing time of tunnels and found further evidence of radio and diurnal effects influencing the core network. With this data in mind, an initial $M/M/\infty$ queue was created to model load occurring at the Gateway GPRS Support Node (GGSN) with simple stationary analysis. This also serves as a basis for future more detailed models.

We think that this investigation and load modeling can lead to better network planning. Being more aware of the control plane provides the necessary tools to identify probable causes for control plane activity. We would also like to expand our evaluations, as there are several angles not investigated so far that could prove worthwhile. This includes

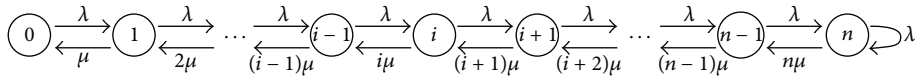



Figure 12: Markov chain model for the tunnel serving process.

an examination of the exact number and size of signaling messages flowing through the core, a more detailed picture of the processing load these messages induce at the GGSN, and an evolved model. Furthermore, a differential analysis of our data compared to a newer dataset (potentially including Long Term Evolution (LTE) access) could really prove worthwhile.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

The work was funded partly by the FTW strategic project "URSA Major." The authors would also like to thank Katharina Salzlechner and Steffen Gebert for their contribution.

## References

[1] J. H. Saltzer, D. P. Reed, and D. D. Clark, "End-to-end arguments in system design," *ACM Transactions on Computer Systems*, vol. 2, no. 4, pp. 277–288, 1984.

[2] S. Corner, "Angry Birds + Android + ads = network overload," 2011, http://www.itwire.com/business-it-news/networking/47823-angry-birds-android-ads-network-overload.

[3] M. Donegan, "Android signaling storm rises in Japan," 2012, http://www.lightreading.com/blog.asp?blog_sectionid=414&doc_id=216929&f_src=lrdailynewsletter.

[4] F. Metzger, S. Gebert, P. Romirer-Maierhofer, K. Salzlechner, A. Rafetseder, and K. Tutschku, "Research report on signaling load and tunnel management in a 3G core network," Research Report FTW-TECHREPORT-121, University of Wuerzburg, Würzburg, Germany, 2012.

[5] 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 (MobiSys '11)*, pp. 321–334, Washington, DC, USA, July 2011.

[6] P. P. C. Lee, T. Bu, and T. Woo, "On the detection of signaling DoS attacks on 3G wireless networks," in *Proceedings of the 26th IEEE International Conference on Computer Communications (INFOCOM '07)*, pp. 1289–1297, Anchorage, Alaska, USA, May 2007.

[7] P. H. J. Perala, A. Barbuzzi, G. Boggia, and K. Pentikousis, "Theory and practice of RRC state transitions in UMTS networks," in *Proceedings of the IEEE GLOBECOM Workshops*, pp. 1–6, Honolulu, Hawaii, USA, November 2009.

[8] C. Schwartz, T. Hoßfeld, F. Lehrieder, and P. Tran-Gia, "Angry apps: the impact of network timer selection on power consumption, signalling load, and web QoE," *Journal of Computer Networks and Communications*, vol. 2013, Article ID 176217, 13 pages, 2013.

[9] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices," in *Proceedings of the ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '11)*, pp. 305–316, San Jose, Calif, USA, June 2011.

[10] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *Proceedings of the IEEE INFOCOM*, pp. 882–890, Shanghai, China, April 2011.

[11] Y. Zhang and A. Årvidsson, "Understanding the characteristics of cellular data traffic," *ACM SIGCOMM Computer Communication Review*, vol. 42, no. 4, pp. 461–466, 2012.

[12] P. Svoboda, F. Ricciato, E. Hasenleithner, and R. Pilz, "Composition of GPRS, UMTS traffic: snapshots from a live network," in *Proceedings of the 4th International Workshop on Internet Performance, Simulation, Monitoring and Measurement (IPS MoMe '06)*, pp. 42–44, Salzburg, Austria, 2006.

[13] X. He, P. P. C. Lee, L. Pan, C. He, and J. C. S. Lui, "A panoramic view of 3G data/control-plane traffic: mobile device perspective," in *Proceedings of the 11th International IFIP TC6 Conference on Networking (IFIP '12)*, pp. 318–330, Prague, Czech Republic, May 2012.

[14] F. Qian, Z. Wang, A. Gerber, Z. M. Mao, S. Sen, and O. Spatschek, "Characterizing radio resource allocation for 3G networks," in *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement (IMC '10)*, pp. 137–150, Melbourne, Australia, 2010.

[15] P. Romirer-Maierhofer, F. Ricciato, and A. Coluccia, "Explorative analysis of one-way delays in a mobile 3G network," in *Proceedings of the 16th IEEE Workshop on Local and Metropolitan Area Networks (LANMAN '08)*, pp. 73–78, Transylvania, Romania, September 2008.

[16] 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.

[17] 3GPP, "General Packet Radio Service (GPRS); Service description; Stage 2. TS 23.060. 3rd Generation Partnership Project (3GPP)," September 2008, http://www.3gpp.org/ftp/Specs/html-info/23060.htm.

[18] 3GPP, "General Packet Radio Service (GPRS); GPRS Tunnelling Protocol (GTP) across the Gn and Gp interface. TS 29.060. 3rd Generation Partnership Project (3GPP)," September 2008, http://www.3gpp.org/ftp/Specs/html-info/29060.htm.

[19] F. Ricciato, "Traffic monitoring and analysis for the optimization of a 3G network," *IEEE Wireless Communications*, vol. 13, no. 6, pp. 42–49, 2006.

[20] 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.

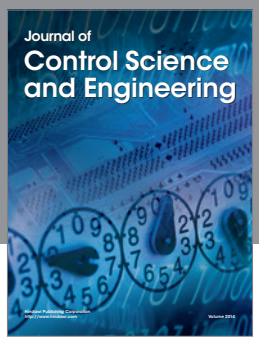
[21] 3GPP, "Numbering, addressing and identification. TS 23.003. 3rd Generation Partnership Project (3GPP)," September 2008, http://www.3gpp.org/ftp/Specs/html-info/23003.htm.

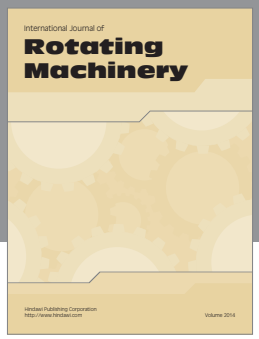
[22] S. F. Donnelly, *High precision timing in passive measurements of data networks [Ph.D. thesis]*, Waikato University, Hamilton, New Zealand, 2002.

[23] C. Mulliner, "Public Research TAC Database," http://www.mulliner.org/tacdb/.

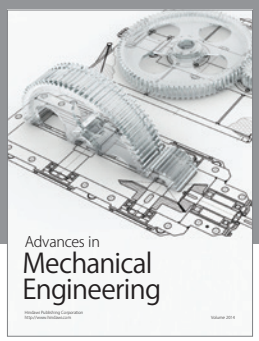
[24] Z. Wang, Z. Qian, Q. Xu, Z. M. Mao, and M. Zhang, "An untold story of middleboxes in cellular networks," in *Proceedings of the ACM SIGCOMM Conference (SIGCOMM '11)*, pp. 374–385, Toronto, Canada, August 2011.

[25] GSM Association, "Fast dormancy best ractises," 2012, http://www.slideshare.net/zahidtg/gsma-fast-dormancy-best-practices.

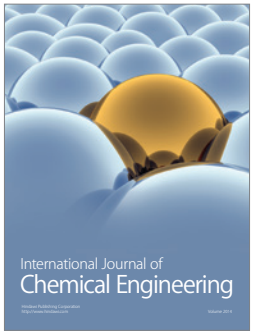



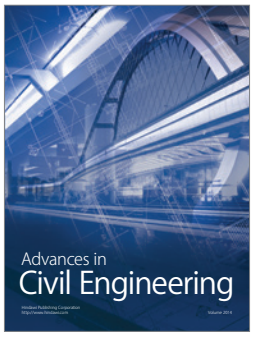


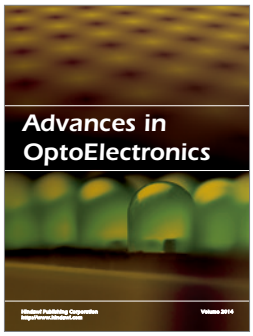




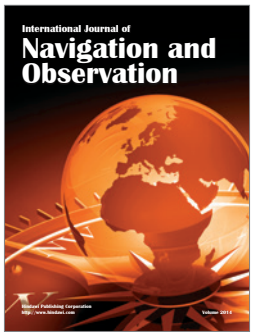


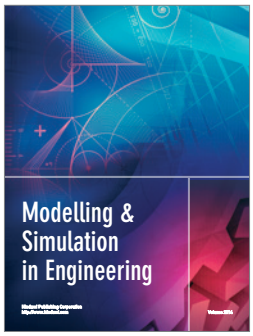




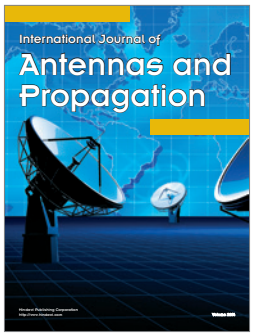





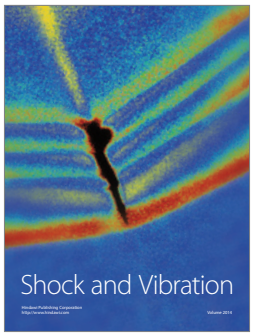










Hindawi

Submit your manuscripts at
http://www.hindawi.com