Characterization of 3G Control-Plane Signaling Overhead from a Data-Plane Perspective

Li Qian† Edmond W. W. Chan† Patrick P. C. Lee‡ Cheng He†
†Noah's Ark Lab, Huawei Technologies, Shenzhen, China
‡Dept of Computer Science & Engineering, The Chinese University of Hong Kong, Hong Kong
{qianli,edmond.chan,hecheng}@huawei.com, pclee@cse.cuhk.edu.hk

ABSTRACT

In 3G networks, when user applications of mobile subscribers send or receive data-plane traffic, control-plane signaling messages will be triggered to initiate or release radio resources for the data transfer. Such signaling messages can increase the processing and transmission overheads of the 3G cellular network infrastructure, and this in turns degrades the performance experience of mobile subscribers. Thus, understanding the signaling overhead of a 3G network becomes critical. In this paper, we conduct the first comprehensive measurement study of the signaling overhead of a city-wide 3G operational network in China. Our main contributions are two-fold. First, based on real cellular traces collected from both data and control planes, we validate that by simply monitoring data-plane packets, we can accurately profile the control-plane signaling overhead due to the initiations of radio resources for data transfer. Second, using data-plane signaling profiling, we characterize the signaling overhead due to common transport protocols and network applications. Our measurement methodology and results presented in this paper would be useful for network operators to better understand how data-plane traffic patterns influence the control-plane signaling overhead of a 3G network.

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design—*Wireless communication*; C.4 [**Performance of Systems**]: Measurement techniques

Keywords

Signaling overhead, 3G networks

1. INTRODUCTION

Third generation (3G) cellular networks have been widely deployed worldwide. With the emergence of smartphones and mobile applications, the performance of operational cellular networks has been challenged. One potential challenge is the control-plane signaling overhead. According to existing 3G standards, signaling messages will be triggered when the user applications of mobile subscribers initiate or release radio resources for data transfer [2]. Different types of user applications (e.g., bulk-data transfer versus interactive applications) have different data transfer characteristics, thereby having different impacts on triggering the signaling messages. Since signaling messages consume radio resources, excessive signaling messages can increase both the processing and transmission overheads within the cellular infrastructure, and hence degrades performance experience of mobile subscribers. From the network operators' perspectives, it is critical to characterize the control-plane signaling overhead of a 3G network due to different types of data-plane traffic so as to devise effective strategies for radio resource provisioning and network planning.

In this paper, we conduct an in-depth measurement study of the signaling overhead of a city-wide 3G operational cellular network in China. At the core of the 3G network, we collect and analyze a 24-hour span of IP data packets and radio resource control (RRC) logs, which correspond to data-plane and control-plane traffic traces, respectively. Our goal is to provide a comprehensive view of how data-plane traffic is correlated with the control-plane signaling overhead of a currently deployed 3G network, from the network operators' perspectives. To our knowledge, this is the first measurement study that quantifies the signaling overhead of a commercially deployed 3G network.

Our contributions are two-fold. We start with considering a data-plane signaling profiling approach that characterizes the 3G signaling overhead by monitoring only raw IP packets in the data plane, without looking into the control-plane information. Our methodology is built on [12, 14], and infers the state transitions for radio resource allocation triggered by data-plane packets. Using our collected cellular traces, we validate that by simply monitoring data-plane packets, we can accurately profile the state transitions due to RRC connection setups, which account for a large proportion of signaling messages. Thus, data-plane signaling profiling simplifies the complexities of collecting and correlating both data-plane and control-plane traces, while still providing highly accurate signaling profiling results.

We further use the data-plane signaling profiling approach to analyze the signaling overhead due to different characteristics of the transport and application layers in the data plane. In each protocol layer, we identify the types of messages that trigger the most signaling messages. In general, we observe that nearly 90% of all state transitions (which trigger signaling messages) are due to small-size data packets with payload less than 200 bytes. We also identify several signaling-prone applications that we observe.

The remaining of the paper proceeds as follows. In Section 2, we review the design of a UMTS network and its radio resource control mechanism. In Section 3, we present and validate our data-plane signaling profiling approach. In Section 4, we report our analysis results on how the signaling load is influenced by the transport-

MSWiM'12, October 21–25, 2012, Paphos, Cyprus.

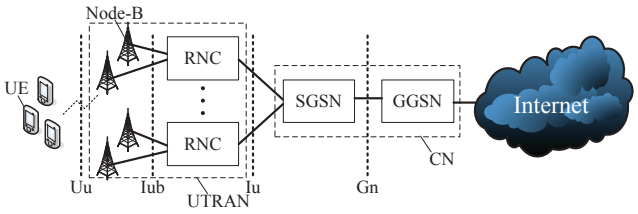


Figure 1: The UMTS network architecture.

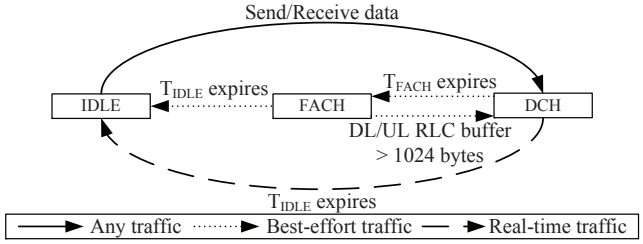


Figure 2: The RRC state machine for the UMTS network considered in this paper.

layer and application-layer behaviors in the data plane. Section 5 reviews related work. Finally, Section 6 concludes the paper.

2. BACKGROUND

2.1 The UMTS network

Universal Mobile Telecommunication System (UMTS) [1] is one of the most popular 3G mobile communication technologies deployed nowadays. It comprises three interacting domains: Core Network (CN), UMTS Terrestrial Radio Access Network (UTRAN), and User Equipment (UE). The CN mainly provides switching, routing, and transit for user traffic and is further divided into circuit switched (CS) and packet switched (PS) domains. In this paper, we focus on the PS domain, which manages IP packet transmissions. Figure 1 shows a typical UMTS network. The UTRAN comprises the base stations (Node-Bs) for serving multiple UEs and the Radio Network Controllers (RNCs). The CN comprises the Serving GPRS Support Node (SGSN) and the Gateway GPRS Support Node (GGSN), each of which communicates with several RNCs for traffic exchange. Moreover, UMTS defines several interfaces, including Gn (between GGSN and SGSN), Iu (between SGSN and RNC), Iub (between RNC and Node-B), and Uu (between Node-B and UE).

2.2 RRC state machine

The Radio Resource Control (RRC) protocol belongs to the control plane of UMTS for system-wide control of communication resources and services. It controls the radio bearer resources for data-plane traffic of each UE. In the protocol, each UE is associated with a *state machine*, which is typically composed of three states: IDLE, CELL_DCH (or DCH), and CELL_FACH (or FACH). Each state corresponds to different levels of radio resource allocation. Initially, when there is no data traffic, a UE is at IDLE and no RRC connection has been established. After an RRC connection is established for data-plane transmission, the UE either operates at DCH for high-speed data transmission through dedicated transport resources, or at FACH for low-speed data transmission in the shared channel to save both radio resources and the UE's energy consumption. Note that in each state transition, the UE and its associated RNC exchange signaling messages through the Iub and Uu interfaces.

Figure 2 depicts the RRC state machine for a commercial UMTS network considered in this paper. When the UE is at IDLE and has uplink (UL)/downlink (DL) data to send/receive, it is always allocated a dedicated transport channel after an RRC connection is set up (i.e., IDLE→DCH). The subsequent state transitions depend on two inactivity timers denoted by T_{FACH} and T_{IDLE}, both of which can be configured by network operators, as well as the service type, which can be specified by the UE in its RRC connection request. For best-effort services (e.g., Web Browsing and Instant Messaging), the RNC resets T_{FACH} to 5s when any UL/DL data traffic is observed, or demotes the state to FACH (i.e., DCH→FACH) when T_{FACH} expires. At FACH, the RNC promotes the state to DCH (i.e., FACH→DCH) again when the Radio Link Control (RLC) buffer size from either the UE (for UL packets) or the RNC (for DL packets) is greater than a threshold of 1024 bytes. However, for real-time services (VoIP and Streaming), DCH is always maintained for the UE to fully utilize radio resources.

Table 1: Numbers of signaling messages generated per transition by each type of RRC state transitions.

RRC state transition	No. of signaling messages
IDLE→DCH	23
DCH→FACH	4
FACH→DCH	10
DCH→IDLE	8
FACH→IDLE	6

To release the RRC connection, the RNC also maintains a T_{IDLE} timer for the UE regardless of the service type. At DCH or FACH, T_{IDLE} is reset to 18s whenever UL/DL data traffic is observed. When T_{IDLE} expires, the RNC demotes the state to IDLE from either DCH (i.e., DCH→IDLE) or FACH (i.e., FACH→IDLE) .

We notice two distinct features from the two RRC state machines reported in [12, 14] compared with ours. First, the transitions from the two state machines do not depend on the service type, thus trading the real-time service performance for improved channel resource utilization among different services. Second, the DCH→IDLE transition never occurs in their state transitions, but instead the RRC connection always switches to FACH before reaching IDLE.

For each state transition, the associated RNC exchanges a few round-trips of signaling messages in the control plane with the UE, SGSN, and Node-B to manage the usage of radio resources allocated for the UE. Table 1 shows the number of signaling messages generated per transition by each type of the RRC state transitions as shown in Figure 2, whose numbers are obtained through internal conversations with our product line engineers. We see that the state promotions (i.e., IDLE→DCH and FACH→DCH) trigger more signaling messages than the state demotions (i.e., DCH→FACH, DCH→IDLE, and FACH→IDLE). Specifically, the IDLE→DCH transition generates the most number of signaling messages. Note that frequent RRC state transitions introduce massive signaling messages, which can overload the processing capacity of the RNC and exhausting radio resources [9].

3. METHODOLOGY VALIDATION

As described in Section 2.2, the state transitions, and hence the signaling messages, are triggered by the data transmissions (either uplink or downlink) for a UE. To understand the root causes of the signaling load of a network, one may need to correlate the control-plane signaling messages and the data-plane packets. Such cor-

relation requires that both traces of signaling messages and data packets be available. In this section, we consider a signaling profiling strategy that uses only data-plane packets to infer the signaling overhead of a network. Using real traffic traces, we validate that data-plane profiling, even without control-plane information, can achieve high inference accuracy in practice. This simplifies the complexities of collecting and correlating the data-plane and control-plane traces.

3.1 Data-Plane Signaling Profiling Overview

We consider a data-plane signaling profiling approach, in which we infer the signaling overhead due to RRC state transitions based on raw IP packet traces captured from the Iu/Gn interfaces. Our profiling approach is built on [12, 14]. Each raw IP packet contains a private IP address that corresponds to a UE. We extract from the traces each *UE session*, which consists of all packets having the same private IP address and packet inter-arrival times less than 60 seconds. We assume that *each private IP address corresponds to a unique UE*, based on the observation that a private IP address is less likely reused by different UEs within a short time period.

For each UE session, we infer a sequence of RRC states based on the inter-arrival times of adjacent IP packets and the known RRC state machine shown in Figure 2. Also, the service type is typically specified in the RRC connection request and is not available in data packets (see Section 2.2). To infer the service type (i.e., real-time or best-effort services) for each IP packet, we use a commercial Deep Packet Inspection (DPI) tool to classify each packet into one of the 14 application types: Database, Email, File Access, Game, Instant Messaging (IM), Network Admin, Network Storage, Peer-to-peer (P2P), Remote Connectivity, Stock, Streaming, Tunneling, Voice over IP (VoIP), and Web Browsing. After identifying the service type, we initiate the required inactivity timers and determine the corresponding state transitions (see Section 2.2). Based on each state transition, we then determine the corresponding number of signaling messages stated in Table 1.

We point out that using data-plane packet traces to infer radio resource usage is not entirely new. Qian *et al.* employ raw IP packet traces collected from two UMTS networks to infer the state machine configurations and energy consumption in each state [12, 14], and propose a client-based profiling tool [13] to perform a similar analysis from the traces collected within a UE. We elaborate the differences of our work and the related studies [12, 13, 14] in Section 5.

3.2 Dataset

It is important that our data-plane signaling profiling approach provides high inference accuracy in practice. In this subsection, we describe the real traffic traces that we use for validating our profiling approach.

In this paper, we analyze 24-hour IP packet traces collected from a commercial UMTS network from a city in China on 1 December 2010. The raw IP packets with full payload are captured (without sampling) from the Iu interface (i.e., the RNC-SGSN interface) and stored in the PCAP format. The network has a total of 16 RNCs deployed in the whole city, yet our analysis mainly focuses on a particular RNC deployed in the urban region. Thus, after the extraction of useful data, we conduct our analysis on 306M IP packets with 682K UE sessions.

During the period, we also collected Performance Call History Record (PCHR) log files from the particular RNC we consider. A PCHR describes the control-plane information about RRC connection setups and releases from a specific RNC for both PS and CS domains. Note that we only use PCHR logs for our validation, while the profiling approach itself does not involve PCHR logs.

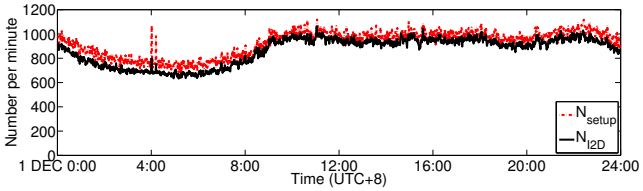


Figure 3: 24-hour time series for the number (per minute) of RRC connection setups (N_{setup}) and the inferred number of IDLE→DCH transitions (N_{I2D}) on 1 December 2010.

3.3 Validation

We now validate the accuracy of our data-plane signaling profiling approach using PCHR logs. We extract all RRC connection setup records in the PS domain from the PCHR log files of the RNC that we consider. Since a single RRC connection setup is triggered by the IDLE→DCH transition of a UE, we can use the setup records to obtain the actual number of IDLE→DCH transitions.

Figure 3 shows the time series of the number of IDLE→DCH transitions (denoted by N_{I2D}) inferred from the data-plane signaling profiling approach and the number of RNC connection setups (denoted by N_{setup}) obtained from the PCHR logs on 1 December 2010. The x-axis is presented in the local UTC+8 China time. Each data point is measured every minute. The figure reveals strong correlation between the inferred number of IDLE→DCH transitions and the actual number of connection setups. Using the profiling approach, we can identify the diurnal pattern of the RRC connection setups. The number of IDLE→DCH transitions drops to below 700 per minute during the midnight, and consistently stays above 800 per minute during the office hour and the evening.

Figure 4 depicts the CCDF of the relative difference between N_{I2D} and N_{setup} using our one-day dataset. The relative difference is computed as $(N_{I2D} - N_{setup})/N_{setup}$. As shown in the figure, the number of RRC connection setups and the inferred number of IDLE→DCH transitions are *very similar*, where more than 80% of the estimates have the relative differences less than 10%. We observe that our profiling approach tends to underestimate the number of IDLE→DCH transitions. The missing IDLE→DCH transitions are likely due to an early RRC connection release by a UE before the 18-second idle time. In this case, the successor of the UE's IP address can immediately transmit a packet through a new RRC connection. Therefore, our signaling profiling approach will miss one IDLE→DCH transition if the inter-arrival time between the last packet and the first packet in the two adjacent RRC connections is less than the dormancy time. Note that the inference accuracy can be improved by using International Mobile Subscriber Identification Number (IMSI) information to distinguish separate RRC connections.

4. SIGNALING PROFILING ANALYSIS

We apply our signaling profiling approach to study the transport-layer and application-layer characteristics for the IP packets after state transitions and hence the signaling overhead on the UMTS network that we consider. We also analyze the root causes of our profiling results, and discuss the possible strategies that can mitigate the signaling load. Our measurement methodology and results provide insights into the impact of data-plane traffic on the signaling load of a commercial UMTS network, and we expect that such insights are of interest to network operators for better network planning.

Our analysis is applied to the data-plane traces (i.e., raw IP pack-

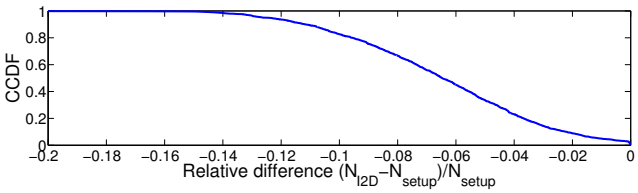


Figure 4: CCDF of the relative difference between the numbers of the RRC connection setups and the inferred IDLE→DCH transitions.

Table 2: Percentage of signaling messages contributed by the state transitions in our dataset

RRC state transition	Percentage of signaling messages
IDLE→DCH	42.84%
DCH→FACH	14.12%
FACH→DCH	25.03%
DCH→IDLE	11.86%
FACH→IDLE	6.15%

ets) gleaned from a particular RNC in our one-day dataset. We collect the *first* IP packets observed right after one of the three state transitions: IDLE→DCH (or I→D in short), FACH→DCH (F→D), and DCH→FACH (D→F). As discussed in Section 2.2, the state promotions I→D and F→D are triggered due to packet transmissions. The state demotion D→F, on the other hand, can be regarded as the result of the delayed packet transmission. By considering the first packets after D→F, we can also characterize how frequently a particular type of application initiates the data transmission in a (low-speed) shared channel. Therefore, analyzing the first packet right after each such transition allows us to infer the root causes of the signaling overheads induced by these state transitions.

On the other hand, we ignore the state demotions DCH→IDLE and FACH→IDLE because they are resulted from inactivity timer expiries and not associated with any packet transmission.

Table 2 shows the percentage of signaling messages contributed by each type of state transitions in our dataset, obtained by counting the number of occurrences of each type of state transitions in the dataset (based on our data-plane signaling profiling) and the number of signaling messages introduced per transition (based on Table 1). We note that the DCH→IDLE and FACH→IDLE transitions altogether contribute only 18% of the total number of signaling messages in the dataset, so their influences on the overall analysis are expected to be insignificant.

We analyze the transport-layer protocol and network application type for each of the collected IP packets. The analysis only focuses on the TCP and UDP packets, both of which contribute the majority of the data traffic. We extract the transport-layer information, such as source/destination ports, TCP flags, and payload sizes from their transport-layer headers. For the application-layer analysis, we use a commercial DPI tool to classify each TCP/UDP packet into one of the 14 application types (see Section 3.2). In addition, we classify whether the packet is observed from the uplink (i.e., from UE to remote destination) or downlink. For an uplink (or downlink) packet, its source (or destination) IP address from the GTPU header is equal to the RNC's IP address. Thus, we can use this information to distinguish between uplink and downlink packets.

We address the limitations of our analysis. First, our dataset is collected in December 2010. Given the emergence of mobile Internet access, the scale and pattern of our dataset may not entirely

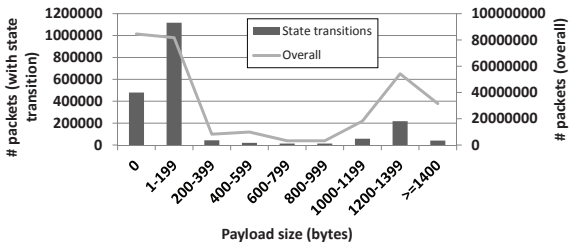
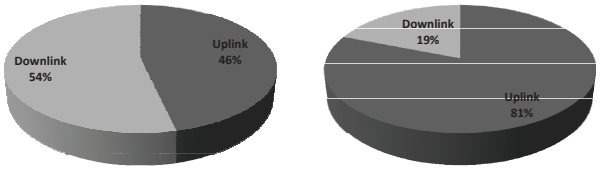


Figure 5: Distributions of payload sizes of overall traffic and the packets inducing state transitions.



(a) Overall traffic.

(b) Only traffic triggering state transitions.

Figure 6: Distributions of uplink and downlink packet numbers.

reflect the real usage of today's 3G networks. Second, although we manually validate some of the results of our DPI tool, we cannot comprehensively verify its correctness as its source code is not available. Third, our dataset, while representing one-day traffic of a city-wide network, may be less representative for some networks in national scales. Nevertheless, our analysis is built on the existing 3G UMTS standard, and we expect that the same analysis methodology is applicable for many 3G operational networks. Our analysis results can provide guidelines for network planning and future analysis.

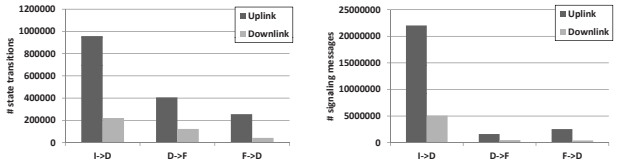
4.1 State Transition Behavior: An Overview

We first provide a general overview of how the state transition behavior is influenced by the packet size and the directions of packet transmissions.

Effect of TCP/UDP payload size. Figure 5 plots the TCP/UDP payload size distributions of overall traffic and the packets inducing state transitions. We observe that 88.6% of the state transitions are induced by small TCP/UDP packets with payload size less than 200 bytes, which altogether contribute 56.4% of the total number of TCP/UDP packets observed from the RNC. Moreover, the packets with zero payload size contribute 23.9% of the state transitions, and all of them are due to *TCP control messages* (e.g., pure ACKs, SYNs, RSTs, FINs) that do not carry any payload.

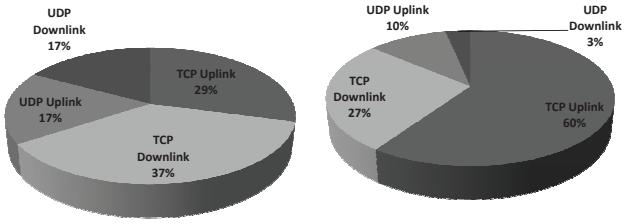
Uplink versus downlink packets. Figure 6(a) plots the distributions of the total numbers of uplink and downlink TCP/UDP packets observed, and Figure 6(b) shows the distributions for only the packets that induce state transitions. As shown in Figure 6(a), the numbers of uplink and downlink packets are fairly close, such that the downlink packets contribute 8% more than the uplink packets. However, Figure 6(b) reveals that more than 80% of the state transitions are triggered by the uplink packets.

We also plot the breakdowns of the TCP/UDP packet-induced state transitions and their corresponding numbers of signaling messages, as shown in Figures 7(a)–7(b). We observe from Figure 7(a) that I→D contributes the most state transitions for both uplink and downlink directions (i.e., 59.1% and 57.3%, respectively), whereas



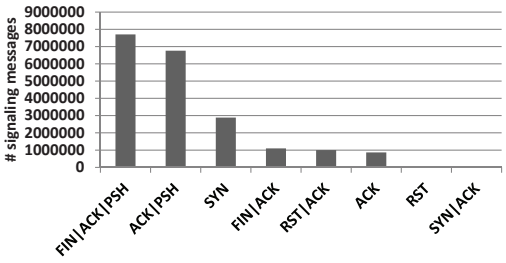
(a) Number of state transitions. (b) Number of signaling messages.

Figure 7: State transition breakdowns of uplink and downlink traffic.

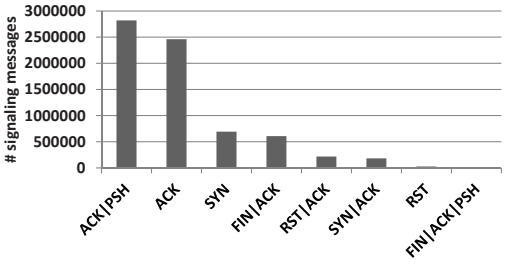


(a) Overall traffic. (b) Only traffic triggering state transitions.

Figure 8: Distributions of the TCP and UDP traffic from uplink and downlink directions.



(a) Uplink traffic.



(b) Downlink traffic.

Figure 9: Distribution of signaling messages triggered by different types of TCP packets.

F→D contributes the least (i.e., 15.8% and 10.9%, respectively). Since each I→D triggers at least twice more signaling messages than other state transitions (see Table 1), Figure 7(b) shows that the contributions of signaling messages of I→D are even more dominant, with 84.1% and 84.8%) in the uplink and downlink directions, respectively. The number of signaling messages due to F→D in the uplink direction is higher than the number due to D→F, mainly because each F→D generates four more signaling messages than each D→F.

Summary of observations. Most signaling messages are triggered by small-size packets (with no more than 200 bytes), including TCP control messages that have zero payload size. Also, most signaling messages triggered by uplink packets initiated by UEs. In the following discussion, we report our transport-layer analysis and application-layer analysis to explain the uneven distributions of the signaling messages observed in the uplink and downlink directions. We also characterize the state transitions induced by the TCP control messages.

4.2 Transport-layer Analysis

In this subsection, we look into the transport layer and analyze how TCP and UDP packets each contribute to the signaling load. In particular, we focus on signaling load due to different types of TCP packets.

TCP versus UDP. Figure 8(a) shows the distributions of the numbers of all TCP and UDP packets, and Figure 8(b) shows the distributions for only the packets that induce state transitions. Figure 8(a) shows that TCP packets contribute a higher proportion than UDP packets (66% versus 34%). We observe more downlink TCP packets (37%) than uplink TCP packets (29%). We can also observe an even distribution of the uplink/downlink UDP traffic (17% each). On the other hand, TCP traffic is more likely to trigger state transitions than UDP traffic. Figure 8(b) reveals that a total of 87% of the packets that trigger state transitions are due to TCP.

TCP flag analysis. Since most state transitions are induced by TCP packets, we investigate different types of TCP packets to identify the root cause of generating state transitions. We classify the TCP packets by their flags set.

Figure 9 shows the number of signaling messages generated by different types of uplink and downlink TCP packets. Both Figures 9(a) and 9(b) show only the top eight types of TCP packets (in descending order) that trigger the most signaling messages. Figure 9(a) shows that among all uplink packets, a significant proportion of signaling messages are due to the FIN/ACK/PSH packets (37.9%) and ACK/PSH packets (33.2%). Figure 9(b) shows that among all downlink packets, ACK/PSH packets generate the most signaling messages (40.2%). Since the signaling messages are mostly triggered by uplink packets, the downlink ACK/PSH packets contribute only 10.3% of the overall signaling messages. Another important observation is that uplink packets carrying control flags SYN, FIN, or RST contribute a significant proportion (46.5%) of signaling messages, while the packets from downlink only introduce a small proportion (6.3%) of messages.

We study how each type of TCP packets (classified by TCP flags and direction) triggers different types of state transitions. Here, we only consider the top eight types of uplink/downlink TCP packets that trigger the most signaling messages. Figure 10 shows the proportions of different types of state transitions for the top eight types of TCP packets (arranged in descending order). For each of the top-four TCP packets FIN/ACK/PSH (UL), ACK/PSH (UL), SYN (UL), and ACK/PSH (DL), more than 50% of the state transitions are due to I→D. Their I→D transitions altogether contribute 45.7% of the state transitions (corresponding to 64.1% of the signaling messages) among all TCP packets that trigger state transitions. Note that almost all (at least 99%) FIN/ACK/PSH (UL), ACK/PSH (UL), and ACK/PSH (DL) packets contain TCP payload, while all of the SYN (UL) packets are pure control messages with zero payload size.

Recall from Figure 9 that RST packets trigger relatively fewer frequent state transitions compared to other types of TCP packets. It is expected because in normal situation a TCP sender immedi-

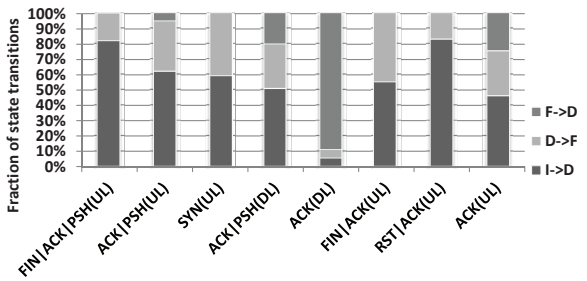


Figure 10: Breakdown of the state transitions triggered by top eight TCP packets (in descending order) that trigger the largest number of signaling messages.

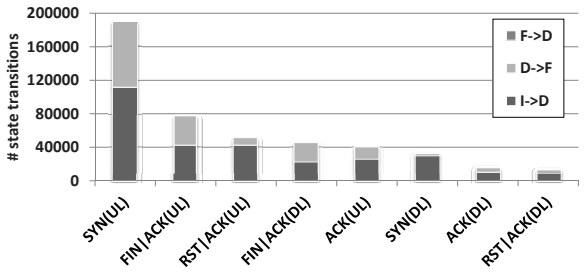


Figure 11: Breakdown of the state transitions triggered by top eight TCP packets without TCP payload (in descending order) that trigger the largest number of signaling messages.

ately dispatches a pure TCP RST (without payload) when it detects incorrect information from an incoming TCP packets [11]. Therefore, TCP RST packets should less likely trigger I→D transitions after an idle period. However, Figure 10 shows that 82.7% of the state transitions induced by RST/ACK (UL) are due to the I→D. One plausible explanation is that some network applications exploit TCP RST messages (instead of TCP FIN messages) to abort TCP connections.

We also note that downlink packets trigger the majority (86.9%) of the F→D transitions. 99.7% of the transitions are due to ACK (DL) and ACK/PSH (DL) packets, and all of them are pure data packets (i.e., ACK with payload). Uplink packets, on the other hand, are more likely to encounter the D→F transitions (76.9%).

TCP control messages. As shown in Figure 5, a substantial proportion (23.9%) of state transitions are due to the TCP control messages with zero payload size. Figure 11 plots the state transition breakdowns for the top TCP control packets that trigger the most signaling messages. The top three of them are SYN, FIN/ACK, and RST/ACK, all of which are in the uplink directions. Due to small packet sizes, it is unlikely for the control packets to induce the F→D transition.

Summary of observations. We observe that most signaling messages are triggered by TCP packets. Many of them are uplink packets initiated by UEs and also carry flags SYN, FIN, or RST, thus accounting for the 64.3% of the I→D transitions inferred from the uplink traffic. We also note that TCP control messages (with null payload) trigger a significant number of I→D transitions.

4.3 Application-layer Analysis

In this subsection, we study how different types of applications (classified by our commercial DPI tool as stated in Section 3.2) contribute to the signaling load. We also seek to differentiate between the signaling-prone and signaling-averse applications.

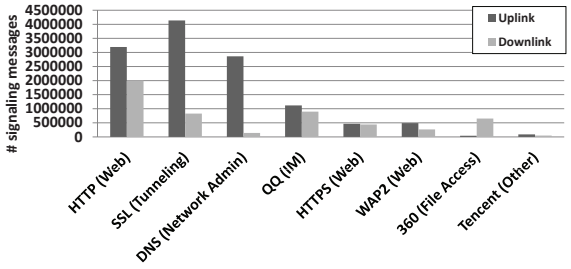


Figure 12: Signaling message distribution of the top eight network applications inducing the largest number of signaling messages (in descending order).

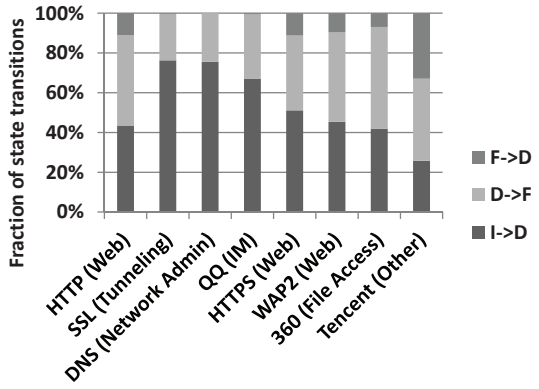
Signaling loads triggered by applications. Figure 12 shows the top eight network applications that trigger the most signaling messages (arranged in descending order). Except for 360 (File Access), all applications generate more signaling messages in uplink packets. In the uplink direction, SSL introduces the most signaling messages (20.3%); in the downlink direction, HTTP introduces the most signaling messages (28.5%).

Figure 13 provides the state transition breakdowns for the top eight applications. We observe that in the uplink direction, most state transitions are I→D. In particular, for SSL, DNS, and QQ, the I→D account for 50% of state transitions. In the downlink direction, DNS's and 360's downlink packets are prone to triggering I→D transitions (with at least 50%), while HTTP, SSL, HTTPS, and WAP2 are prone to triggering F→D transitions (67.7%, 85%, 48.1%, and 62.1%, respectively). These applications introduce significant numbers of ACK or ACK/PSH packets in the downlink, thus accounting for the majority of F→D transitions (see Section 4.2). All the top eight applications are best-effort services; at least 20% of the transitions observed from each such application are due to D→F.

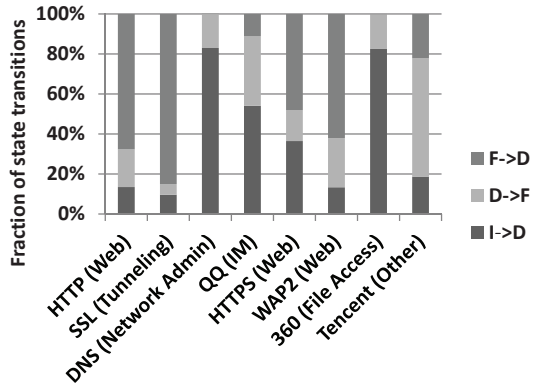
Signaling-prone and signaling-averse applications. In the above application-layer analysis, the contribution of signaling messages of each application depends on the amount of traffic being generated. We now attempt to identify how the uplink and downlink traffic patterns of an application affect the signaling load, independent of the amount of traffic being generated.

Here, we look into the *signaling density* of each application. Consider the uplink traffic of an application. Let $N_{packets}$ be the total number of uplink packets and N_{trans} be the total number of induced state transitions. We then compute the signaling density, defined by $N_{trans}/N_{packets}$, for each type of network applications whose $N_{packets}$ is greater than 0.1% of the overall uplink traffic observed from the RNC. This avoids having a very large ratio due to the very small number of packets observed. Thus, an application is *signaling-prone* if $N_{trans}/N_{packets}$ is high, or *signaling-averse* if $N_{trans}/N_{packets}$ is low. We also repeat the above procedures with the downlink traffic to identify signaling-prone and signaling-averse applications.

Figures 14 and 15 report the signaling density and the fraction of traffic for the top signaling-prone and signaling-averse applications determined by their uplink and downlink traffic, respectively. As shown in Figure 14, several applications (e.g., SSL and QQ) are signaling-prone for both uplink and downlink directions, while some of them, especially network administration applications like SSDP, induce signaling overheads only from a particular direction of traffic. On the other hand, Figure 15 shows that bulk-transfer applications such as streaming, peer-to-peer, and file access applica-

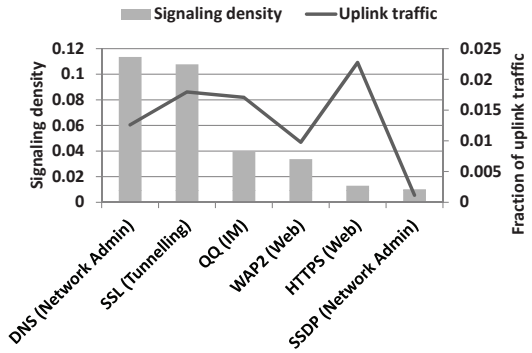


(a) Uplink traffic.

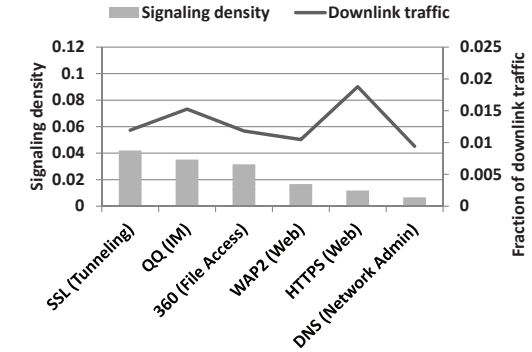


(b) Downlink traffic.

Figure 13: Breakdown of the state transitions triggered by uplink and downlink traffic for the top eight network applications.



(a) Uplink traffic.



(b) Downlink traffic.

Figure 14: Signaling density and traffic distributions of the top signaling-prone network applications in the uplink and downlink directions.

tions are less likely to introduce signaling messages in both uplink and downlink directions.

Summary of observations. We observe that the signaling-prone applications that generate the most signaling messages are all interactive applications, including Web, Tunneling, Network Admin, and IM. Such interactive applications generate messages with a large inter-packet time, so they generate signaling messages if the inter-packet time is larger than the idle period. The protocol messages of these applications are generally of small size, so this explains why most signaling messages are triggered by small-size packets (see Section 4.1).

4.4 Discussion

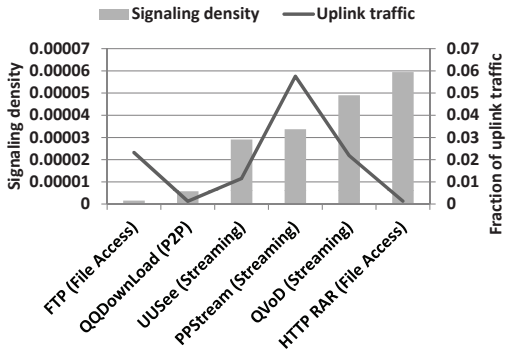
In this section, we analyze data traffic traces of a 3G operational network, and identify different types of transport-layer and application-layer data traffic that trigger most signaling messages. We observe that most signaling messages are attributed to the I→D transitions, i.e., when the RRC connection is activated from the idle state. To mitigate the signaling load (from a network operator's perspective), one possible solution is to keep the RRC connection of a mobile active for a longer period of time, such as by polling the mobile or increasing the inactivity timeout value. The trade-off is that it increases the energy consumption of the mobile [12, 14]. Another possible solution is to optimize the signaling mechanism, although it may also contradict the existing 3G standard. In future work, we explore strategies that can mitigate the signaling load, based on the transport-layer and application-layer traffic patterns.

5. RELATED WORK

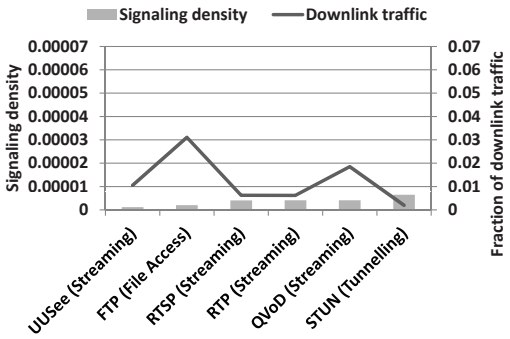
There have been various measurement studies on cellular data networks. Existing studies mainly analyze the *data-plane* performance by looking into the cellular traces at the network core. Examples of analysis include: similarities with IP wireline traffic [15], performance of TCP flows [8], user browsing behavior [7], spatial/temporal traffic variations of mobile subscribers and network resource usage [10], as well as traffic patterns of smartphone applications [16].

Recent measurement studies on cellular data networks also look into the *control-plane* performance, mainly from the security perspective. Lee *et al.* [9] show that low-volume data traffic with specific patterns can overload the signaling plane of a Radio Network Controller (RNC) in a 3G network. Zhao *et al.* [17] consider a similar type of signaling attack, but focus on the IP Multimedia Subsystem (IMS) services deployed in a 3G network. Both studies [9, 17] focus on the signaling load on a 3G network similar to ours. However, their studies are based on simulations, while ours is based on real traces of a 3G operational network.

The closest related works to ours are [12, 14], which infer the radio resource control (RRC) state machine using data packets collected at the core of a 3G network. Qian *et al.* [12] uses the RRC state machine to characterize the energy consumption of a mobile in each RRC state. Their follow-up work [14] proposes a fast-dormancy-based solution to proactively notify the 3G network of the release of the radio resources so as to achieve better power saving. On the other hand, our work has several key differences from [12, 14]. First, we focus on the signaling load of a cellular network and hence we are more concerned with the state transitions, each



(a) Uplink traffic.



(b) Downlink traffic.

Figure 15: Signaling density and traffic distributions of the top signaling-averse network applications in the uplink and downlink directions.

of which triggers a different load of signaling messages. On the other hand, [12, 14] focus on the energy consumption of a mobile handset, which is mainly attributed to the times spent on non-idle states (i.e., FACH and DCH). Second, we use a more comprehensive dataset which includes all IP-level packets with full payload, while [12, 14] focus on TCP traffic only. We also use deep packet inspection (DPI) to identify the signaling loads of different types of applications. Finally, and most notably, we use the control-plane RRC logs to validate the accuracy of our signaling inference, while the validation is not considered in [12, 14].

Besides using cellular traces, some studies [3, 4, 6, 13] collect traces directly within mobile handsets. In particular, [13] uses the traces of a mobile handset to infer the RRC state machine. Our work conducts measurements from a network operator's perspective, such that we have full access to the traffic traces at the core.

Our prior work [5] conducts a measurement study of the data/control-plane performance of different types of mobile terminals in a 3G network. This work complements our prior work by quantifying the signaling overhead of a 3G network and studying how the signaling load is influenced by the characteristics of the transport and application layers.

6. CONCLUSIONS

To our knowledge, we provide the first measurement study of the impact of raw data packets (in the data plane) on the signaling load (in the control plane) of a commercial city-wide 3G network in China, based on real traffic traces collected at the network core. We consider a data-plane signaling profiling approach that uses only raw IP packet traces to infer the signaling load, and most notably, we validate with real traces that such a profiling approach achieves high inference accuracy. We proceed to use the data-plane signaling profiling approach to analyze the interactions between raw data packets and the signaling load. In particular, we identify the types of data packets, in both transport and application layers, that contribute most to the overall signaling load.

7. ACKNOWLEDGMENTS

The work of Patrick Lee is supported in part by grant GRF CUHK 413711 from the Research Grant Council of Hong Kong.

8. REFERENCES

[1] 3GPP. UMTS. `http://www.3gpp.org/article/umts`.

[2] 3GPP. UTRAN Functions, Examples on Signaling Procedures (Release 1999), Jun 2002. TR 25.931 v3.7.0.

[3] H. Falaki, D. Lymberopoulos, R. Mahajan, S. Kandula, and D. Estrin. A First Look at Traffic on Smartphones. In *Proc. of ACM IMC*, Nov 2010.

[4] H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin. Diversity in Smartphone Usage. In *Proc. of ACM MobiSys*, Jun 2010.

[5] 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 *Proc. of IFIP/TC6 Networking*, 2012.

[6] J. Huang, Q. Xu, B. Tiwana, Z. M. Mao, M. Zhang, and P. Bahl. Anatomizing Application Performance Differences on Smartphones. In *Proc. of ACM MobiSys*, 2010.

[7] R. Keralapura, A. Nucci, Z.-L. Zhang, and L. Gao. Profiling Users in a 3G Network Using Hourglass Co-Clustering. In *Proc. of ACM MobiCom*, 2010.

[8] J. Kilpi and P. Lassila. Micro- and macroscopic analysis of RTT variability in GPRS and UMTS networks. In *Proc. of NETWORKING*, 2006.

[9] P. P. C. Lee, T. Bu, and T. Woo. On the Detection of Signaling DoS attacks on 3G/WiMax Wireless Networks. *Computer Networks*, 53(15):2601–2616, Oct 2009.

[10] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das. Understanding Traffic Dynamics in Cellular Data Networks. In *Proc. of IEEE INFOCOM*, 2011.

[11] J. Postel. Transmission Control Protocol, Sep 1981. RFC 793.

[12] F. Qian, Z. Wang, A. Gerber, Z. Mao, S. Sen, and O. Spatscheck. Characterizing Radio Resource Allocation for 3G Networks. In *Proc. of ACM/USENIX IMC*, 2010.

[13] F. Qian, Z. Wang, A. Gerber, Z. Mao, S. Sen, and O. Spatscheck. Profiling Resource Usage for Mobile Applications: a Cross-Layer Approach. In *Proc. of ACM/USENIX MobiSys*, 2011.

[14] F. Qian, Z. Wang, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck. TOP: Tail Optimization Protocol for Cellular Radio Resource Allocation. In *Proc. of IEEE ICNP*, 2010.

[15] J. Ridoux, A. Nucci, and D. Veitch. Seeing the Difference in IP Traffic: Wireless versus Wireline. In *Proc. of IEEE INFOCOM*, 2006.

[16] Q. Xu, J. Erman, A. Gerber, Z. M. Mao, J. Pang, and S. Venkataraman. Identifying Diverse Usage Behaviors of Smartphone Apps. In *Proc. of ACM IMC*, Nov 2011.

[17] B. Zhao, C. Chi, W. Gao, S. Zhu, and G. Cao. A Chain Reaction DoS Attack on 3G Networks: Analysis and Defenses. In *Proc. of IEEE INFOCOM*, 2009.