Control-Plane Protocol Interactions in Cellular Networks

Guan-Hua Tu[†*], Yuanjie Li[†*], Chunyi Peng[‡], Chi-Yu Li[†], Hongyi Wang[†], Songwu Lu[†]
[†]University of California, Los Angeles [‡]The Ohio State University, Columbus
[†]{ghtu, yuanjie.li, lichiyu, hywang, slu}@cs.ucla.edu [‡]chunyi@cse.ohio-state.edu

ABSTRACT

Control-plane protocols are complex in cellular networks. They communicate with one another along three dimensions of cross layers, cross (circuit-switched and packet-switched) domains, and cross (3G and 4G) systems. In this work, we propose signaling diagnosis tools and uncover six instances of problematic interactions. Such control-plane issues span both design defects in the 3GPP standards and operational slips by carriers. They are more damaging than data-plane failures. In the worst-case scenario, users may be out of service in 4G, or get stuck in 3G. We deduce root causes, propose solutions, and summarize learned lessons.

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design—*Wireless Communication*; C.4 [**Performance of Systems**]: *Design Studies, Modeling Techniques*

Keywords

Cellular networks; control-plane; protocol verification

1. INTRODUCTION

The cellular network serves as a large-scale wireless infrastructure to support mobile data and voice services. A salient feature of its design has been its control-plane protocols. Compared with the Internet, these components provide more complex signaling functions. They follow the layered protocol architecture (see Figure 1 for an illustration), and run at both the network infrastructure and the end device. Together, they provide control utilities vital to 3G/4G networks, including mobility support, radio resource control, session management for data and voice, *etc*..

In this paper, we examine protocol interactions in cellular networks. We focus on a set of critical components on the control plane (see Table 2 for the list). Our goal is to uncover problems during inter-protocol communications. Although each signaling protocol may be well designed individually, proper interactions among them in the networked environment are not guaranteed.

There are two challenges. First, compared with the Internet, cellular networks are still closed systems. Signaling exchanges are not readily accessible from carriers, nor from devices during normal operations. Second, patterns of inter-protocol communication on the control plane are much richer than their Internet counterparts. In addition to the inter-layer case, they exhibit in both cross-domain and cross-system scenarios in cellular networks. Since both data and carrier-grade voice are indispensable services, both packet switching (PS) and circuit switching (CS) are used. Signaling protocols thus regulate both PS and CS domains. Moreover, Inter-system switching between 3G and 4G is also common due to hybrid deployment, user mobility, or CSFB (CS Fallback)-based calls. Signaling protocols consequently need to work cross 3G and 4G systems. In a nutshell, interactions among control-plane protocols are common in 3G/4G systems. They span three dimensions: between layers of the protocol stack (cross-layer[1]), between CS and PS domains (cross-domain), and between 3G and 4G systems (cross-system).

In this work, we devise *CNetVerifier*, a tool to analyze all such cases (§3). Our tool adapts model-checking methods with cellular-specific heuristics. It further instruments the device to collect protocol traces for validation. We apply the tool and delve into all above three dimensions. Our study yields interesting findings (§4). We show two classes of problematic interactions among signaling protocols. They are exemplified using six concrete instances, spanning cross-layer, cross-domain, and cross-system dimensions (see Table 1). In the first class (§5), we show that some inter-protocol communications are necessary yet troublesome. The necessity of signaling synergy is partly driven by the requirement for carrier-grade voice support, partly by inter-system switching in hybrid 3G/4G deployments, and partly by mobility management. However, interactions among signaling protocols are not always designed and operated right: (S1) a user device is temporarily out of service because its vital context in 4G is shared but not well protected (being deleted after inter-system switching); (S2) Users are denied network access right after being accepted because higher-layer protocols make unrealistic assumptions on lower layers; (S3) 4G users get stuck in 3G because inconsistent policies are used for CS and PS domains in 3G and 4G. The second class (§6) concerns independent operations by protocols. We discover that, some are unnecessarily coupled and have unexpected consequence: (S4) outgoing calls are delayed for unjustified location updates because cross-layer actions are "improperly" correlated and prioritized; (S5) PS data sessions suffer from rate reduction ($51\% - 96\%$ drop observed) when traffic in both domains shares the same channel; (S6) User devices are out of service when the failure is propagated to another system. We validate most instances with traces collected from our tool when running tests over two US carriers. We further conduct a two-week user study to assess their real-world impact (§7). We propose and evaluate solutions that help to resolve above issues (§8 and §9).

2. BACKGROUND

The cellular network architecture consists of base stations (BSes) and a core network. The BSes provide radio access to user devices (*e.g.*, phones), whereas the core network connects them to the wired Internet or the public telephony network. Figure 1 illustrates the network architecture and main protocols for both 3G and 4G.

The 4G LTE network offers PS data service only. It has three core elements: (1) MME (Mobility Management Entity) to manage

*The first two authors contribute equally to this work. The correspondence faculty author is C. Peng.

[1]We use inter-layer and cross-layer interchangeably in this paper, for a slight abuse of definition.

SIGCOMM'14, August 17–22, 2014, Chicago, IL, USA.

http://dx.doi.org/10.1145/2619239.2626302.

Category	Problems	Type	Protocols	Dimension	Root Causes
Necessary but problematic cooperations	**S1:** User device is temporarily "*out-of-service*" during 3G→4G switching.	Design	SM/ESM, GMM/EMM	Cross-system	States are shared but unprotected between 3G and 4G; States are deleted during inter-system switching (§5.1).
	S2: User device is temporarily "*out-of-service*" during the attach procedure.	Design	EMM, 4G-RRC	Cross-layer	MME assumes reliable transfer of signals by RRC; RRC cannot ensure it (§5.2).
	S3: User device gets stuck in 3G.	Design	3G-RRC, CM, SM	Cross-domain; Cross-system	RRC state change policy is inconsistent for inter-system switching (§5.3).
Independent but coupled operations	**S4:** Outgoing call/Internet access is delayed.	Design	CM/MM, SM/GMM	Cross-layer	Location update does not need to be, but is served with higher priority than outgoing call/data requests (§6.1).
	S5: PS rate declines (*e.g.*, 96.1% in OP-II) during ongoing CS service.	Operation	3G-RRC, CM, SM	Cross-domain	3G-RRC configures the shared channel with a single modulation scheme for both data and voice (§6.2).
	S6: User device is temporarily "*out-of-service*" after 3G→4G switching.	Operation	MM, EMM	Cross-system	Information and action on location update failure in 3G are exposed to 4G (§6.3).

Table 1: Finding summary.

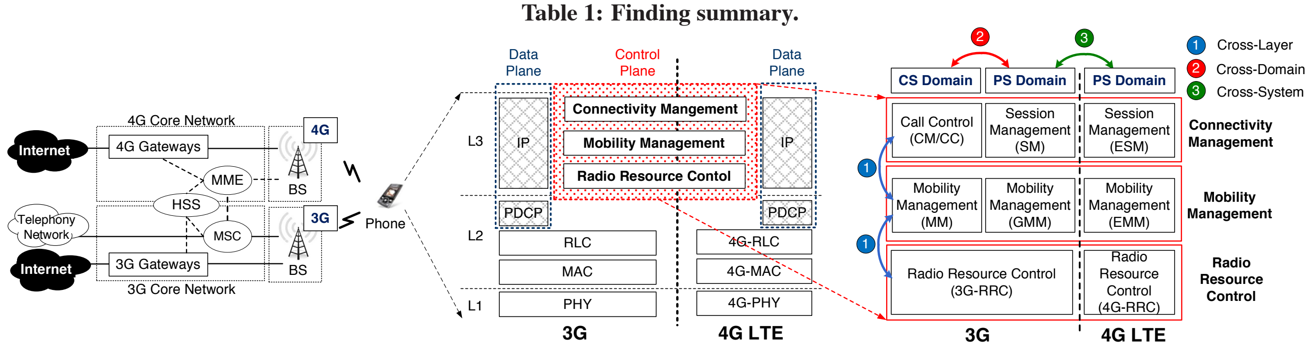


Figure 1: 4G/3G network architecture and control-protocol interactions in three dimensions.

user mobility (*e.g.*, location update or paging), (2) 4G gateways that route PS packets between the Internet and the 4G BSes, and (3) HSS (Home Subscriber Server), which stores user subscription information. In contrast, the 3G network supports both CS and PS services. Its core network consists of: (1) MSC (Mobile Switching Center), which pages and establishes CS services (*i.e.*, voice calls) with mobile devices, (2) 3G Gateways, which forward PS data packets, and (3) HSS, which is similar to its counterpart in 4G.

Similar to the Internet, cellular network protocols have adopted the layered structure. The protocol family spans both data and control planes. The data plane is responsible for actual data and voice transfer. The control plane provides a variety of signaling functions to facilitate the data-plane operations. Specifically, three major functions are provisioned at three sub-layers: (1) Connectivity Management (CM), which is responsible for creating and mandating voice calls and data sessions; (2) Mobility Management (MM), which provides location update and mobility support for call/data sessions; (3) Radio Resource Control (RRC), which controls radio resources and helps to route signaling messages.

We next introduce major procedures in cellular networks.

Attach/detach cellular networks. The mobile device must *attach* to the cellular networks before using any cellular network service[2] (*e.g.*, data or voice service). It happens when the device powers on. The attach procedure is mandated by Mobility Management control protocols (*i.e.*, MM, GMM and EMM) running on mobile devices, 3G MSC, 3G Gateways and 4G MME, respectively. Once it completes, the mobile device is "*registered*" until being detached. The *detach* procedure can be triggered either by the device (*e.g.*, the phone powers off) or the network (*e.g.*, under resource constraints). Once detached, the device enters the "*deregistered*" (*i.e.*, "out-of-service") state and cannot access any cellular service.

Data and voice services. Both are essential services offered by cellular networks. To enable any *data* service, the mobile device has to first establish a bearer with the core network. This procedure is done via "EPS Bearer activation" in 4G or "PDP Context activation" in 3G, which is mandated by Evolved Session Management (ESM in 4G) or Session Management (SM in 3G). Once it succeeds, the core network assigns an IP address, reserves resources to meet QoS requirements and establishes the routing path for the device. In fact, the information vital to data sessions (*e.g.*, IP address and QoS parameters) is stored at both the device and the 3G/4G gateways via the 3G PDP (or 4G EPS bearer) context.

In 3G, the voice service is supported via CS and handled by the Call Control (CC) protocol at the mobile device and MSC. In 4G, the voice service is designed to run over PS via Voice-over-LTE (VoLTE) [2]. However, due to the high deployment cost and complexity of VoLTE, most 4G operators adopt another voice solution, Circuit-Switched Fallback (CSFB), which switches 4G users to legacy 3G and accesses CS voice service in 3G [5].

Radio resource control (RRC). RRC is responsible for controlling radio resources between the device and the BS. An established RRC connection is the prerequisite for any communication (data, voice or signaling) between the device and the core network. A RRC state machine is used for this purpose. Two states of `IDLE` and `CONNECTED` denote whether the RRC connection has been established or not. For goals of optimization and energy efficiency, 3G and 4G also offer multiple connected sub-states. Specifically, 3G uses `FACH` and `DCH`. The former supports low-rate communication with less radio resource and power consumption, whereas the latter consumes more but sends packets at higher speed. In contrast, 4G supports three modes of continuous reception, short and long discontinuous reception.

Mobility management. The cellular network supports two types of mobility: (1) intra-system handover, where the user stays within 3G or 4G only and updates its location during roaming. It is done by one of the following procedures: location area update via MSC (3G CS), routing area update via 3G Gateways (3G PS) or tracking area update via MME (4G); (2) inter-system switch, where the user device switches between 3G and 4G. Once the switch succeeds, the device updates its location to the new serving network via the above procedure. For signaling protocols, mobility support is realized through MM, GMM, and EMM in 3G CS, 3G PS and 4G PS (see Figure 1), respectively. The underlying radio access switch (*e.g.*, radio channel setup/teardown) is handled by 3G/4G RRC .

[2] The only exception is to make emergency calls.

3. METHODOLOGY

We develop *CNetVerifier*, a tool that conducts two-phase protocol diagnosis, as shown in Figure 2. It helps to uncover two types of issues: *(i) design problems* originated from the 3GPP standards, and *(ii) operational slips* originated from the carrier practice.

3.1 CNetVerifier Overview

CNetVerifier takes a two-phase approach. During the *screening* phase, *CNetVerifier* first explores *possible* logical design defects in control-plane protocols via model-checking techniques, and produces counterexamples due to design defects. Once they are found, we move to the *validation* phase. For each counterexample, we set up the corresponding experimental scenario and conduct measurements over operational networks for validation.

We use the two-phrase approach since both phases are necessary. The issues discovered during the first phase are implementation- and measurement-independent ones, since they come from the 3GPP design standards. Moreover, its outputs (*i.e.*, these counterexamples) offer us hints to configure the experiments to validate possible design problems. The second phase alone may not uncover all problematic issues since it is measurement dependent. This phase is needed for validating the design problems and studying their impact. Moreover, it helps to identify operational slips or implementation bugs. For example, S5 and S6 are found during the S3's validation experiments.

Before elaborating techniques for each phase, we rush to point out several downsides of *CNetVerifier*. First, it focuses on the control-plane protocol interactions, thus simplifying data-plane operations (*e.g.*, ignoring packet communication time and call durations). Second, the defined properties are from the user's perspective. It may not uncover all issues at base stations and in the core network which operators are interested in. Third, using random sampling for usage scenarios, some parameter-sensitive defects may not be exposed. The impact could be alleviated by increasing sampling rates. Fourth, due to limited access to cellular networks, some findings may not be validated by experiments. For example, S2 is discovered by protocol screening but not observed through phone-based experiments. We cannot confirm whether it rarely happens or it is not a real defect. Finally, we mainly conduct experiments according to those counterexamples reported during the *screening* phase. Not all operational slips may be identified.

3.2 Domain-Specific Protocol Screening

During protocol screening, we discover the issues originated from cellular network design. To this end, we develop a cellular-specific model-checking tool, which is written in Spin [12]. It works as follows. First, we model signaling protocol interactions, and define cellular-oriented properties. Second, given these inputs, *CNetVerifier* checks whether a set of desired properties are satisfied. It thus generates a counterexample for each concrete instance of property violation, which indicates a possible design defect. To make the above idea work in the cellular context, we address three domain-specific issues: (1) How to model cellular networks? (2) How to define the desired properties? (3) How to check the property given the cellular network model?

3.2.1 Modeling

Our modeling effort covers both parts of 3G/4G protocol stacks and usage scenarios. The protocol interactions occur between protocols in the stack, and are driven by usage scenarios.

Modeling 3G/4G protocol stacks. The modeling of cellular protocols is derived from the 3GPP standards [3,6–8], which specify the operations for each protocol. Table 2 lists the studied cellular protocols, including PS/CS services, mobility management and radio resource control. We model each cellular protocol as two Finite State Machines (FSMs), one running at the user device and the other operating in the specific network element (for instance, CM/MM, SM/GMM, ESM/EMM are operated at MSC, 3G Gateways and MME, respectively).

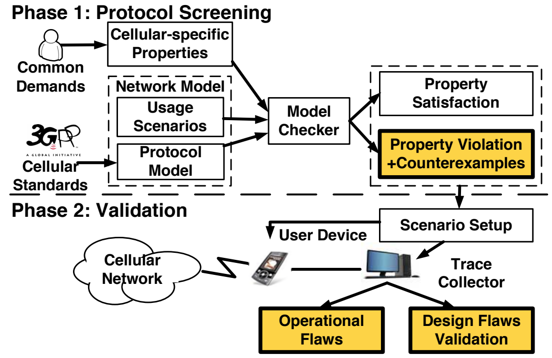


Figure 2: ***CNetVerifier*** **Overview**

Function	Name	System	Net. Element	Standard	Description
PS/CS	CM/CC	3G	MSC	TS24.008	CS Connectivity Management
	SM	3G	3G Gateways	TS24.008	PS Session Management
	ESM	4G	MME	TS24.301	4G Session Management
Mobility	MM	3G	MSC	TS24.008	CS Mobility Management
	GMM	3G	3G Gateways	TS24.008	PS Mobility Management
	EMM	4G	MME	TS24.301	4G Mobility Management
Radio	3G-RRC	3G	3G BS	TS25.331	Radio Resource Control
	4G-RRC	4G	4G BS	TS36.331	Radio Resource Control

Table 2: Studied protocols on network elements and devices.

Modeling usage scenarios. Modeling usage scenarios is more challenging. They are not formally defined by the 3GPP standards, and largely depend on user demands and operation policies. Ideally, we should test all combinations of usage scenarios, so that all possible design defects can be found. However, some usage scenarios may have unlimited choices. Enumeration is thus deemed unrealistic. Consequently, we take the random sampling approach. We assign each usage scenario with certain probability, and randomly sample all possible usage scenarios. Specifically, for scenarios with limited options (*e.g.*, device switch on/off, all types of accept/reject requests, all inter-system switch techniques), we enumerate all possible combinations. For scenarios with unbounded options (*e.g.*, user mobility at various speed, traffic arrival patterns of PS/CS services), we implement a run-time signal generator that randomly activates these options at any time. Last, each customizable parameter is initialized with a random value. By increasing the sampling rate, we expect that more defects can be revealed. Specifically, we model user demands and operator responses as follows.

◦ *User demands* In our model, the phone device uses at most one network at a time, and cannot concurrently access both 3G and 4G networks. This is the default practice for most smartphones in reality. Once the device powers on, it randomly attaches to 3G or 4G. Afterwards, a run-time signal generator randomly creates user-specific events, such as starting voice or data service, location change or user-initiated detach (*i.e.*, switch off). These events thus trigger relevant protocol entities at the device to respond accordingly and further activate procedures towards the network.

◦ *Operator responses* Upon receiving a user request, the network accepts or rejects it. We equally test with all the possibilities, including the reject with various error causes. For example, more than 30 error causes are defined in the 4G attach procedure [8]. In the meantime, the run-time signal generator randomly produces network-specific events, *e.g.*, inter-system switch and network-oriented detach. Similarly, corresponding procedures towards the user device are triggered. Note that all options for network-specific events are stipulated by the standards and will be enumerated in our model. More details will be given later.

3.2.2 Defining Desirable Properties

In this work, we seek to check those problematic protocol interactions that incur user-perceived problems. The properties to be checked represent the services offered to users. Thus, we define three cellular-oriented properties: (1) *PacketService_OK*: Packet data services should be always available once device attached to 3G/4G, unless being explicitly deactivated. (2) *CallService_OK*: Call services should also be always available. In particular, each call request should not be rejected or delayed without any explicit user operation (*e.g.*, hanging up at the originating device). (3) *MM_OK*: inter-system mobility support should be offered upon request. For example, a 3G↔4G switch request should be served if both 3G/4G are available. We consider inter-system mobility only because intra-system mobility is seamlessly supported in practice. Note that *PacketService_OK* and *CallService_OK* represent the expected behaviors for network services, while *MM_OK* is for mobility support. In *CNetVerifier*, these properties act as logical constraints on the PS/CS/mobility states.

3.2.3 Property Checking

We perform the formal model checking procedure. First, the model checker creates the entire state space by interleaving all FSMs for each individual protocol. With the constraints of three properties, some states will be marked with "error." Then we run the depth-first algorithm to explore the state transitions from the initial state (*i.e.*, the device attempting to attach to 3G/4G networks) under various usage scenarios. Once an error state is hit, a counterexample is generated for the property violation. The model checker finally generates all counterexamples and their violated properties for further experimental validation.

3.3 Phone-based Experimental Validation

Given counterexamples for design defects, the validation phase needs to conduct experiments, collect protocol traces from real networks and compare them with the anticipated operations. The main challenge is trace collection. The core cellular network is operated as a black box, so it is not easy to obtain protocol traces from cellular network operators. Therefore, we seek to retrieve protocol traces from user devices. Fortunately, most cellular modem vendors (*e.g.*, QualComm or Mediatek) allow for developers to power on the debugging mode and obtain protocol traces[3]. Based on this, we collect five types of information: (1) timestamp of the trace item using the format of hh:mm:ss.ms(millisecond), (2) trace type (*e.g.*, STATE), (3) network system (*e.g.*, 3G or 4G), (4) the module generating the traces (*e.g.*, MM or CM/CC), and (5) the basic trace description (*e.g.*, a call is established).

To facilitate PS and CS signaling exchanges, we further devise automatic test tools on the phone. One is to automatically dial out, answer and terminate an incoming voice call. The other is to keep turning on and off data services. We use Speedtest [1] to measure the uplink and downlink speed of the Internet access on the phone. Each experiment has 10 runs unless explicitly specified.

We conduct experiments over two major US operators, denoted as OP-I and OP-II, for privacy concerns. They together serve more than 140M subscribers. We use five smartphone models that support dual 3G and 4G LTE operations: HTC One, LG Optimus G, Samsung Galaxy S4 and Note 2, and Apple iPhone5S. They cover both Android and iOS. All phones are used in all validation experiments. The experimental settings are constructed based on the counterexamples from the screening phase. We also test with common use scenarios to explore whether any operational slip is observed to break three properties in practice.

4. OVERVIEW OF FINDINGS

We uncover signaling interaction problems in both design and operations through *CNetVerifier*. We examine standards specification to identify design issues, and collect protocol traces to infer improper operational practice. Our findings are summarized in Table 1. They are grouped into two classes. The first class, *necessary yet problematic cooperations*, refers to the protocol interactions that are required but misbehave. The second class, *independent yet unnecessarily coupled operations*, refers to the protocol interactions that are not necessary but indeed occur and result in negative impact. The troubling inter-protocol signaling each leads to functional incorrectness or performance penalty. Not all the issues are operational slips, so they cannot be fully fixed by simply updating their implementations. For design problems, 3GPP standards should be revised to address them. Specifically, we first identify four instances S1-S4 in the screening phrase and then uncover two more operational issues S5 and S6 in the validation phrase. In fact, other issues are revealed by *CNetVerifier*, but they are not reported here because they do not belong to problematic protocol interactions. Both classes of issues are found in all three dimensions.

• **Cross-layer** Protocols in the upper-layer and low-layer directly interact with each other via the interfaces between them. Two representative instances are found in this category. In both cases, the principle of protocol layering is not properly honored. In the first case (§5.2), the low-layer RRC protocol fails to offer reliable and in-sequence signal delivery required by the upper-layer EMM protocol. EMM thus should have implemented its own end-to-end mechanism but does not. Subsequently, the signaling exchange between the device and the network can be lost or delayed, triggering wrong reactions from EMM. It denies user's network access right after accepting the access request. In the second case (§6.1), CM/SM and MM/GMM protocols, running on different layers in 3G, should act on outgoing call/data requests and location updates independently and concurrently. However, they prioritize location updates over call/data requests. The head of line blocking is experienced, and the outgoing calls and data are unnecessarily delayed.

• **Cross-domain** In cross-domain protocol interactions, protocol variants are developed for different domains and indirectly coupled over the common lower-layer protocols (*e.g.*, RRC). The cross-domain category also has two cases. In principle, the CS-domain voice and the PS-domain data have distinct requirements. Data prefers high throughput whereas voice values timely delivery. They should be treated differentially. However, in both cases, identical operations are performed on traffic from both domains. In the first case (§5.3), RRC keeps its state for the aggregated CS and PS data traffic. When the CS traffic terminates, the PS data may get stuck in 3G without returning to 4G networks. In the second case (§6.2), carriers use RRC to assign PS and CS sessions on a shared channel, using a single modulation scheme for both voice and data. The PS data rate may drop significantly over the shared channel.

• **Cross-system** Cross-system interactions occur with an 3G↔4G switch. Two instances are further uncovered in this category. In this scenario, both systems may be motivated to share or even act on certain state information. On one hand, correct information should be properly protected and shared during the cross-system operations. This is exemplified by the first case (§5.1). To enable data services, the user and the network must keep the PDP context in 3G and the EPS bearer context in 4G. However, such states are not well protected during inter-system switching. 3G

[3] For example, both QXDM (http://www.qualcomm.com/qxdm) and XCAL-Mobile (http://www.accuver.com) support this mode.

may delete the PDP context, and then the 4G network cannot recover its EPS bearer context. The user device is thus out of service in 4G after the inter-system handover. In the second case (§6.1), 3G and 4G share information on location update failures. The actions on such failures should be confined between 3G and 4G networks. However, 4G takes action on the user device when handling failure signals from 3G. The user consequently loses its network access.

In following §5 and §6, we elaborate on each problematic case. Given each instance, we describe its concrete procedure, deduce its root cause, validate and assess its negative impact over US carriers.

5. IMPROPER COOPERATION

We describe three instances S1-S3 that exhibit troubling interactions in cross-system, cross-layer, and cross-domain settings.

5.1 Unprotected Shared Context in 3G/4G

The first is on *cross-system* signaling interactions between 3G and 4G. When the user device switches from 4G to 3G during mobility or CSFB calls, the data service is indeed migrated accordingly. However, under certain conditions, when the user switches back to the 4G network (*e.g.*, after completing a CSFB call or roaming back to a 4G BS), the device might be temporarily *out of service*. Our experiments validate its existence, and show that this out-of-service status may last from several to tens of seconds in operational networks. It is also quite common in reality. The root cause lies in improper cross-system interactions, and the involved protocols are SM/GMM in 3G and ESM/EMM in 4G, running at two signaling layers of session control and mobility support. These protocols should interact, because they need to support seamless PS data sessions when user devices switch between 3G and 4G. They thus share contexts in 3G and 4G. However, 4G mandates such shared states but 3G may have deleted them, thus causing state recovery failure after successful inter-system handover.

5.1.1 Inter-System Switch

The inter-system switch is commonly observed between 3G and 4G in practice. It occurs in three popular usage settings. First, in hybrid 3G/4G deployment, the mobile user leaves the coverage of current system, enters the cell of another system, and then roams back to the old system. Second, a user makes a CSFB-based call in 4G LTE networks, which triggers two handoffs, *i.e.*, one from 4G to 3G to start the voice call in 3G, and one from 3G to 4G after the call completes. Third, carriers may initiate such switching for users for load balancing or better resource availability. In case PS data access is enabled (when the mobile data network is ON), a 3G↔4G information migration will be performed accordingly. Note that, critical information and states are stored in PDP or EPS bearer context in 3G or 4G before the switching. To ensure smooth migration, the PDP context in 3G and the EPS bearer context in 4G are translated and kept consistent. For example, the IP address, *etc*.. remains the same before and after the switching.

Figure 3 shows how signaling protocols interact during 4G→3G switching[4] [4]. There are three steps. First, 4G RRC at the device receives the command from the 4G base station, disconnects the RRC connection between the device and the base station, and informs EMM. Second, 3G RRC at the device connects to the 3G base station using the information carried in the above command. It informs MM and GMM of such an inter-system switching for both CS and PS domains. MM and GMM subsequently initiate the location update procedure in both 3G CS and PS domains. If any data service was initiated when the device was in 4G, the gateways and MME (in Figure 1) collaborate to transfer the 4G EPS bearer

[4]The scenario shown here is "*RRC connection release with redirect*", which is a typical inter-system switching mechanism.

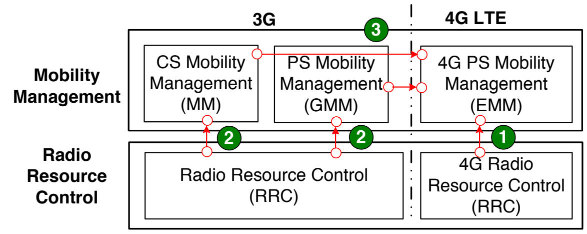


Figure 3: The 4G→3G inter-system switching flow.

context into the 3G PDP context during the location update procedure. After the conversion, the resources reserved for the 4G EPS bearer will be released. Third, MM/GMM in 3G informs EMM in 4G regarding the successful switching. The procedure for 3G→4G switching is similar. The 3G PDP context is migrated to the 4G EPS bearer context during the location update performed in 4G.

5.1.2 Issues and Root Causes

In the instance S1, our tool reports that the above protocols violate the property of *PacketService_OK*. We find that the user becomes *out-of-service* after an inter-system switching.

The scenario is as follows. The user device is initially in 4G and has its EPS bearer context activated. It then switches to 3G in one of the three usage scenarios. The EPS bearer context is subsequently deleted from 4G to release resource reservation. While in 3G, the PDP context can also be deactivated for various reasons (listed in Table 3). However, when later switching back to 4G, the device cannot register to the 4G network, since 4G only supports PS services and EPS bearer context is required. It detaches itself and becomes out of service in 4G. We next understand the root cause and the impact in three aspects.

We first see why the PDP context is deleted in 3G. The EPS bearer context or the PDP context is essential to enabling PS services. Since 4G only supports PS, its EPS bearer context is *mandatory* for data service and signaling exchange. Whenever it cannot be constructed, no service access is available based on the 4G standards [8]. On the other hand, the PDP context in 3G is allowed to be deactivated. It is *not mandatory* in 3G. Since 3G supports both CS and PS, a user can still use the CS voice service without the PDP context. Deactivation of the PDP context is common in 3G. Both the network and the user device can initiate it. It can also be triggered by various reasons (listed in Table 3).

We next look into whether it is a serious issue and how bad its negative impact is. Note that most smartphones do not support dual radios for both 3G and 4G. Each phone thus access one network at any time. Once being detached by 4G, the device has access to neither 4G nor 3G. This can last a few seconds. Of course, the device may immediately seek to re-register to 4G. It leaves the "*out-of-service*" state once registration succeeds. Otherwise, it keeps trying until the maximum retry count is reached. When all retries fail, the device may start to try 3G.

We finally see whether the above problem can be eliminated. The issue can be fully addressed since it stems from a design defect. First, the 3G PDP context does not need to be deactivated in all cases. Therefore, the 4G EPS bearer context can be re-constructed and the device obtains data access after switching from 3G to 4G. For example, the reason "QoS not accepted" in Table 3 states that the QoS cannot be satisfied at the user device. If so, the PDP context can be kept while changing to a lower QoS policy at the phone. The factor "Incompatible PDP context" implies that the active PDP context is not compatible for all PS services, *e.g.*, MMS and Internet. The PDP context can also be modified rather than being deleted. The cause "Regular deactivation" is triggered by the user (*e.g.*, when turning off the mobile data) or by the network. The

Originator	Cause
User device	Insufficient resources
User device	**QoS not accepted**
User device/Network	Low layer failures
User device/Network	**Regular deactivation**
Network	**Incompatible PDP context**
Network	Operator determined barring

Table 3: PDP context deactivation causes.

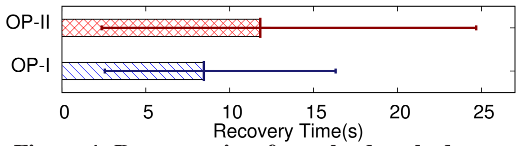


Figure 4: Recovery time from the detached event.

PDP context can also be kept until the switching to 4G succeeds. Second, even the PDP context has to be deactivated in 3G for compelling reasons, the user device can still avoid out-of-service after the inter-system switching. The reason is that, now the user device is still in registered state in 4G, it can reactivate a EPS bearer rather than being detached. This way, the device recovers from the PDP context deactivation.

5.1.3 Experimental Validation

We next conduct experiments to validate and assess the above issue. We run tests to switch phones between 3G and 4G networks and collect protocol traces at the phone. The switching is done through two methods: (1) by CSFB call, and (2) by driving back and forth between two areas covered by 3G and 4G networks. We verify the instance in both OP-I and OP-II in our tested phones. When the device switches to 3G, the PDP context is deactivated by the network. After migrating back to 4G, the phone is detached by 4G due to *"No EPS Bearer Context Activated"* error.

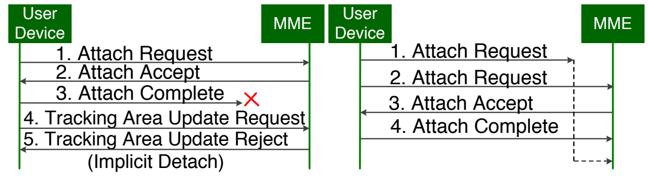
We also observe the same issue when users disable cellular data services or switch to WiFi networks. For most smartphones, they will disable the mobile data service whenever a local WiFi network is accessible. While staying in 3G, some (here, HTC One and LG Optimus G) deactivate all PDP contexts. As a result, when users later switch to 4G, they become out of service for the same error.

We further observe an implementation issue that is complementary to S1. The tested phone may stay in the *out-of-service* state longer than expected. When no PDP context is found during switching to 4G, the phone does not detach immediately by following the 3GPP standards. Instead, it initiates the attach procedure until receiving the message of location update reject from networks. Note that it is not designed in 3GPP standards but observed in our tested phones. Figure 4 plots the median, minimum and maximum recovery time measured on Samsung S4 over more than 50 runs in both carriers. The recovery time is the one from the time when the tracking area update reject is received to the time when re-attach succeeds. We see that the device takes 2.4s to 24.7s to complete the attach procedure. Similar results are observed at other phones (median gap < 0.5s). It is because the re-attach is mainly controlled by operators. The phone is unreachable (*i.e.*, out of service) during the recovery time.

Insight 1: *For the contexts shared between different systems, the actions and policies shall be consistent across systems. Otherwise, cross-system issues may arise.*

5.2 Out-of-Sequenced Signaling in Inter-Protocol Communications

The instance S2 appears during cross-layer protocol interactions in 4G networks. The two involved protocols are EMM and RRC. We find that, the user device may temporarily be "*out-of-service*" and lose 4G access. It is induced by the improper action taken by



(a) Lost signal (b) Duplicate signal

Figure 5: Device is detached by lost/duplicate signals.

EMM when communicating with RRC. The EMM protocol relies on RRC to transfer signals, but assumes reliable, in-sequence signaling messages. The underlying RRC protocol does not provide it. Even worse, the design of EMM does not anticipate any lost or delayed signaling exchange. This leads to unexpected consequence. The user is detached from 4G right after successful attach.

5.2.1 Issues and Root Causes

We find that the above protocol interaction violates the property of *PacketService_OK*. The device enters the "*deregistered*" state (*i.e.*, out of service in 4G), after receiving error signals of either attach reject or location update reject. There are two cases.

○ *Lost signaling messages*. The first case happens when the attach request message is lost. Figure 5(a) plots the signaling sequence during the attach procedure. Initially, EMM at the device sends an attach request to MME in the core network (Step 1), which replies an attach accept (Step 2). The device establishes the EPS bearer , and responds to MME with an attach complete signal (Step 3). However, this signal may be lost when invoking the RRC protocol for transmission to the base station, which further relays it to MME. According to the standards [7], RRC does not always ensure reliable delivery and the signal can be lost (*e.g.*, over the air). Since MME does not receive the attach complete message, inconsistent EMM states exist between the device and MME.

On the user side, he believes the attach procedure succeeds, while MME does not think so. Once the tracking area update (*i.e.*, location update in 4G) is triggered , the problem worsens. During this operation, the user sends the tracking area update request to MME (Step 5). However, upon receiving it, the EMM protocol at MME does not process it since it believes the attach procedure has not completed yet. EMM thus rejects it with error type "*implicitly detach*" and deregisters (*i.e.*, detaches) the device from 4G, which subsequently deletes the EPS bearer context. When receiving this reject message, the user device has to detach itself from the network after the prior attach success.

○ *Duplicate signaling messages*. The second case is observed when duplicate attach requests are received at MME (shown in Figure 5(b)). After sending the attach request (Step 1) through BS1, the mobile user roams to BS2. However, BS1 is under heavy load and defers the delivery of this signal to MME. Since it does not receive the reply message on time, the device retransmits the request signal (Step 2) via BS2 and receives the attach accept from MME. This completes the attach procedure at both the device and MME. However, the duplicated attach request finally arrives at MME via BS1. Given this duplicate signal, standards [8] stipulate that the EPS bearer context is deleted and MME processes the duplicate attach request. Two outcomes are possible. One is that the duplicate request is rejected. The device becomes "*out-of-service*". The other is that it is accepted. The EPS bearer has to be re-constructed, and packet service is unavailable during the transition.

The EMM protocol at MME seems to have valid reasons to take above actions. Whenever it observes incomplete attach (in the first case), EMM has no reason to retain the EPS bearer context for the device. When receiving a new attach request at the registered state for the device, EMM has to reprocess it. Otherwise, it may lead to inconsistent states (*i.e.*, registered or deregistered) at MME and the user device. EMM indeed needs to reprocess the request to resolve inconsistency in other settings. Assume that the device is suddenly out of battery and cannot notify MME. MME still keeps the device in the registered state, thus leading to inconsistency between the device and MME. When the device later powers on after recharge and sends attach request to MME, EMM should process it to recover consistency.

There are two causes rooted in improper cross-layer interaction. First, EMM protocol itself is not prepared for out-of-sequenced signaling exchange. It makes the assumption that the underlying protocols ensure reliable, in-sequence signal delivery. Its design does not consider cases of lost and duplicate signals. Second, end-to-end (*i.e.*, from the device to MME through intermediate base stations) reliable delivery for signals is not readily ensured. This holds true even when reliable delivery is assured between user device and base station, as well as between the base station and MME. The exception arises during user mobility. Signals can be relayed by two different base stations, and the signals may still lose their original sequencing when arriving at MME.

5.2.2 Experimental Validation

In the experiments, we use three approaches to trigger the attach/reattach procedure in 4G: (1) power on and off the 4G-only devices, (2) manually change the network type between 3G-only and 4G-only on the device, and (3) reuse the experiments conducted in §5.1. To make signals lost in the air, we conduct experiments in the areas with weak signal coverage (*i.e.*, RSSI is below -110dBm).

Our tests indeed show that EMM signaling messages are lost when the radio transmission is bad. However, we do not observe the implicit detach due to lost signals. The most common scenario we observe is that user device keeps retransmitting the attach requests, while no attach accept message is received. It is because cellular networks are still closed systems, we are unable to drop or delay specific EMM signals from 4G base stations/MME to validate this design detect. In the future work, we plan to cooperate with operators to investigate network elements at the validation phase.

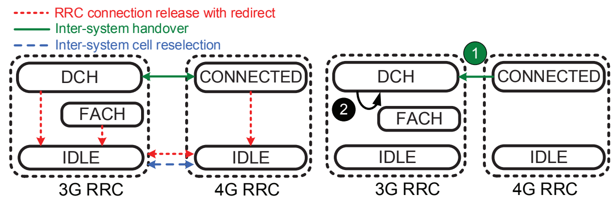
Insight 2: *During cross-layer protocol interactions, the key functionality of upper layer protocols should not merely rely on the non-always-guaranteed features in lower layer protocols. Otherwise, they are operating at the risk of failures.*

5.3 Inconsistent Cross-Domain/Cross-System Protocol State Transition

The third instance S3 is both cross-domain (between 3G CS and 3G PS) and cross-system (between 3G and 4G). We find that, a 4G user device may get stuck in 3G, thus losing its 4G connectivity and high-speed access, after completing a CSFB voice call. This occurs when the device still carries a high-rate data session, regardless of whether the user is roaming or not. Note that this is against the design of CSFB, which should move the device back to 4G after the call. This scenario complements our recent study [27], which only uncovers similar problems but when the device uses low-rate data service. The root cause lies in inconsistent state transition for the RRC protocol when handling both CS-domain voice and PS-domain data in the process of inter-system switching.

5.3.1 Issues and Root Causes

Both instance S3 and that in [27] violate the property *MM_OK* (*i.e.*, inter-system mobility support). The device thus gets stuck in 3G, and cannot go back to 4G after the CSFB call. It happens when a CSFB call has terminated. Specifically, when making the call, the 4G user switches to 3G but still uses data service in the PS domain. Once the call completes, the device intends to switch back to 4G.



(a) Inter-system switch options (b) High-rate data + CSFB

Figure 6: RRC states in various inter-system switching options.

However, this inter-system switching cannot be activated (the property *MM_OK* is violated). We have two observations. First, there is an ongoing PS data session since the PDP context is active. Second, the 3G RRC state is at either `FACH` or `DCH` (*i.e.*, `CONNECTED`).

The root cause lies in the RRC protocol, which regulates both the CS domain and the PS domain during the inter-system switching between 3G and 4G. Figure 6(a) illustrates RRC transitions in three inter-system switching options. The first option, "RRC connection release with redirect", starts with RRC non-IDLE state and forces an RRC connection release before the inter-system switching. It migrates the device back to 4G but disrupts the ongoing high-rate data session. Second, an inter-system handover is invoked. It supports the direct transition between 3G `DCH` and 4G `CONNECTED`. It mitigates interruption of data session but incurs operation overhead for carriers (*e.g.*, buffering and relaying packets during the handover). The third option is "inter-system cell selection". It works for RRC IDLE state and it is triggered by the mobile device to look for better 3G/4G cells for subsequent switching.

The standard gives the carriers freedom to choose these switching options. However, the state transition for inter-system switching has design defects. Figure 6(b) shows the simplified RRC state transition in this CSFB case. When the CSFB call starts, the RRC state migrates from 4G to 3G `DCH` (Step 1) due to the high-rate data service. When the CSFB call in the 3G CS domain completes, RRC remains at the `DCH` state since the high-rate data is still ongoing. It is stuck in 3G if inter-system cell selection option is selected by operators. We see that the RRC state is determined by both CS-domain voice and PS-domain data. Although PS and CS domains do not interact directly, both domains rely on RRC for control. They share the same RRC state. This shows that, signaling interaction between CS and PS domains is done through the RRC protocol. The cross-domain signaling is needed because CS and PS domains are dependent. As long as the CS-based call is ongoing, data session in the PS domain has to stay in 3G. It may move to 4G only after the call terminates.

Carriers should not be held responsible for the deadlock. They do follow the standards. It is understandable for carriers to use "inter-system cell selection" to switch back to 4G after the CSFB call ends. First, it reduces the network loading to monitor and respond to each CSFB call state, since it is triggered by mobile device. Second, it does not interrupt current data sessions. However, the fundamental problem is that, 3G/4G standards fail to design the bullet-proof RRC protocol, which should handle all cross-domain, cross-system scenarios.

5.3.2 Experimental Validation

We start a 60-min UDP uplink/downlink data session at high rate (200kbps) in both OP-I and OP-II. We make a CSFB call from the LTE phone and hang it upon after the call starts. We confirm that the RRC state at the phone remains at `DCH` after the call hangs up. In OP-I, the phone switches to 4G in a few seconds through the option of *RRC Connection Release and Redirect*. Its data session is disrupted. In OP-II, the device gets stuck in 3G. It is the same as the duration of data sessions (about 60 minutes in our experiments).

No	Scenario	Category
1	Cross location area	Location area updating
2	Periodic location update	Location area updating
3	CSFB call ends	Location area updating
4	Cross routing area	Routing area updating
5	Periodic routing update	Routing area updating
6	Switch to 3G system	Location and routing area updating

Table 4: Scenarios trigger location/routing area update.

Insight 3: *The original well-designed features can become error-prone as new functions are enabled. Design options should be prudently justified, tested and regulated. Otherwise, the desirable benefit may be compromised by various unregulated option choices.*

6. PROBLEMATIC COUPLED ACTIONS

We now report three problematic coupling instances, discuss the root causes, and evaluate their impact on users.

6.1 HOL Blocking for Independent Updates

The instance S4 is on unnecessary coupling between cross-layer protocols in 3G. Both voice and data services may suffer from Head of Line (HOL) Blocking and thus extra latency due to independent, yet unnecessarily prioritized location update at underlaying layers. The involved protocols are CM/MM and SM/GMM for the CS domain and the PS domain, respectively.

6.1.1 Issues and Root Causes

The network needs to know the location of the device. Without it, the network cannot route *incoming* calls to the user. Table 4 lists various usage scenarios that may trigger location update. This update is performed for roaming users, and it is also used for periodic refresh without mobility or after inter-system switching. In 3G CS domain, the *location update* is initiated by MM protocol on user device, and sent to MSC. In 3G PS domain, the location update is performed by GMM via *routing area update*, and 3G gateway is responsible for accepting/rejecting it.

CNetVerifier reports that outgoing CS/PS service requests from the CM/SM layer can be delayed while the MM/GMM layer is doing location/routing area update. In CS, the issue arises when an outgoing call is initiated and CM sends the request[5] to MM. However, the CM service request is delayed (or even rejected based on the standards [6]) when MM is running the location update. Similar results can be observed on the cross-layer interaction of GMM and SM in the PS domain. Note that both the outgoing call request and the location update are initiated by the user device in S4 here.

At first sight, the above decision seems to be plausible. Two requests are waiting to be served. One is the CS/PS service request at CM/SM, while the other is the location update request at MM/GMM. The service request should be deferred and yield to the location update. Without correct location information updated at the network, the device is not reachable by others. Location updates should be processed with high priority.

However, this is not well grounded. Note that the call/data request is *outbound*. The device can always send it out. If this call request is served first, MSC also *implicitly* updates the location for the device as a byproduct of call serving. Therefore, *inbound* services are not affected by whether the location update request or the call request is served first. There is no need to serve the location update request in the expedited manner. Implicit update can be realized without any extra resource. The service requests on upper-layer CM/SM protocols are independent of the location updates at lower-layer MM/GMM. Artificially correlating and prioritizing them incur unnecessary latency to user service requests.

[5]It is used to establish the signaling connection between the device and MSC for call setup.

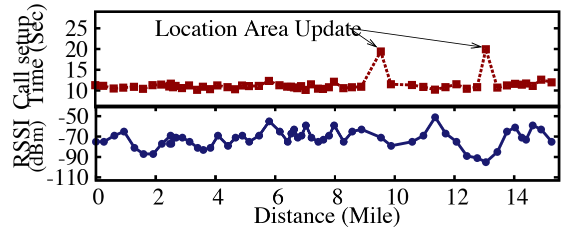
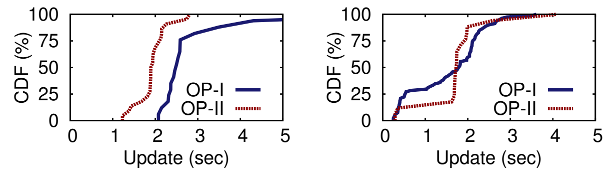


Figure 7: Call setup time and RSSI on Route-1 in OP-I.



(a) Location area update (CS) (b) Routing area update (PS)

Figure 8: CDF of location update durations in OP-I and OP-II.

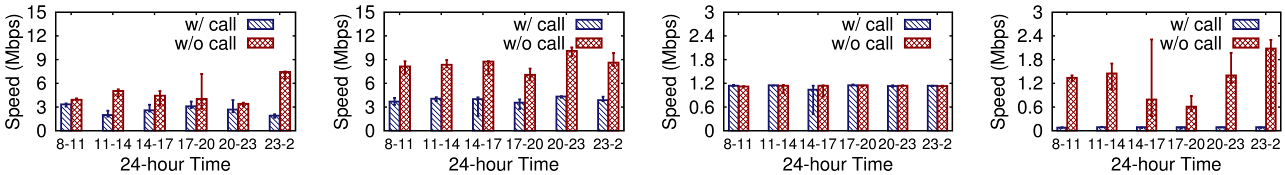
6.1.2 Experimental Validation

Call service. In the experiment, the caller repeatedly dials the callee, and immediately dials again once the callee hangs up. It is done when we drive along two routes: Route-1 (15-mile freeway) and Route-2 (28.3-mile freeway+local), in both OP-I and OP-II. The observed phenomenon is similar between carriers and across test runs. We show results in OP-I only. We indeed see that phones delay the call request until location update is completed. Figure 7 plots the call setup time on Route-1 (*i.e.*, from dialing to connected call) and the measured signal strength (RSSI). The average setup time is around 11.4 seconds, and RSSI varies within the good-signal range [-51dBm, -95dBm]. We observe two location updates at two spots of the route, 9.5 mile (RSSI:-73dBm) and 13.2 mile (RSSI:-87dBm). When the call is initiated during location update, the call setup time increases to 19.7 seconds, about 8.3 seconds longer than the average. Since the measured RSSI is strong, we infer that the extra time is caused by the location update. Figure 8(a) plots the CDF of duration for location area update. In OP-I, all updates take longer than 2 seconds, and the average is about 3 seconds. In OP-II, 72% of routing area updates take 1.2–2.1 seconds, and the average is 1.9 seconds.

We also notice a chain effect for delayed call services. The call requests are delayed for 8.3 seconds, whereas location update takes 4 seconds. It turns out, the extra 4.3 second gap is incurred by MM while it process both cross-layer MM and RRC related commands in the state "*MM-WAIT-FOR-NET-CMD*" [6] after the location update. In this state, all the call requests will be unnecessarily delayed until new commands from network arrive.

Internet data service. In this test, we first turn on the data service and transfer data packets to an Internet server, and then disable the PS service. Our experiments show that, the SM data requests are not immediately processed during the routing area update. Figure 8(b) plots the CDF of duration for routing area update. In OP-I, around 75% of updates take 1-3.6 seconds. In OP-II, 90% of routing area updates take from 1.6 seconds to 4.1 seconds. Therefore, the impact of routing area update in the PS domain is a little bit smaller than location update in the CS domain. This is because GMM does not process RRC related functions, whereas MM has to. However, routing area update is performed more frequently than location update. The user is more likely to experience delayed data service than a deferred outgoing call.

Insight 4: *Some procedures in upper and lower layers seem independent but are coupled by their execution order. Without prudent design, HOL blocking may happen.*



(a) Downlink (OP-I) (b) Downlink (OP-II) (c) Uplink (OP-I) (d) Uplink (OP-II)

Figure 9: Downlink and uplink data speed (maximum, median and minimum) with/without CS calls in both carriers.

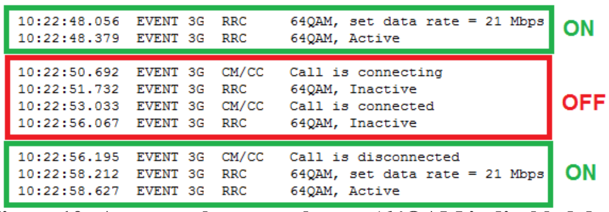


Figure 10: An example protocol trace (64QAM is disabled during CS voice call, OP-I).

6.2 Fate Sharing for Voice and Data

The instance S5 is an operational problem in dual-domain operations. In our experiments, we keep observing fate-sharing on transmission rates between PS and CS domains. When both PS and CS are accessing the 3G network on the phone, the PS data rate decreases significantly, compared with the case of accessing 3GPS only. This is due to improper cross-domain (CS/PS) coupling between PS and CS in 3G. It is implemented by carriers, and does not appear to be a design slip in the standards.

Figure 9 plots the downlink and uplink speed when the PS service is enabled with/without the CS call at different hours of a day. When both services are concurrently enabled, downlink and uplink data rates (expect the uplink rate in OP-I) decrease. It seems reasonable since PS and CS are competing for the shared radio resource. However, given that the best 3G CS voice is 12.2kbps [11], the actual PS data rate degrades beyond expectation (a small or mild drop expected). The downlink decline is up to 3.5-5.8 Mbps, about 73.9% in OP-I and 74.8% in OP-II. The uplink speed drop in OP-II reaches 96.1% (for OP-I, one 51.1% drop observed).

We figure out that, the large rate drop in PS is due to the inappropriate cross-domain channel sharing. In general, CS voice and PS data have different requirements. The CS traffic requires high resilience and low loss to ensure timely delivery and reduce voice message retransmission. It thus prefers the more robust, low-rate modulation scheme (*e.g.*, 16QAM). In contrast, PS traffic prefers high data rate for faster access. It thus prefers high-rate modulation (*e.g.*, 64QAM). Our protocol trace analysis shows that, both carriers configure the phone via the RRC protocol. The phone transfers both CS and PS traffic over the shared channel and apply the same modulation scheme. The modulation scheme is chosen so that the CS traffic is satisfied first, at the cost of PS rate degradation. Figure 10 gives an example trace collected in OP-I. We see that, before the voice call is made, the used modulation scheme is 64QAM, thus offering downlink speed up to 21Mbps. Once the voice call starts, both OP-I and OP-II disable 64QAM. The highest-rate modulation turns 16QAM, thus reducing the theoretical downlink speed to 11Mbps. The user thus suffers from large rate drop in its data service. Certainly, a tradeoff between performance and radio resource control exists. Sending CS and PS traffic over the shared channel may reduce carriers' resource waste [23]. However, it is achieved at the cost of large PS rate decline. The above measurements indicate that current tradeoff is not a good practice from users' perspective.

A different sharing scheme may yield better results. Consider each shared channel used by multiple users allow each to adopt his own modulation scheme; The modulation scheme may change over time due to varying signal strength. Also, one device can use multiple channels. Instead of coupling the CS and PS traffic from the same device on the shared channel, we can cluster PS sessions from multiple devices and let them share the same channel while CS sessions are grouped together and sent over the shared channel using the same modulation scheme. An alternative approach is to allow CS and PS to adopt their own modulation scheme. This way, diverse requirements of CS and PS traffic can both be met.

Insight 5: *When two domains have different goals and properties, their services should be decoupled as possible. Otherwise at least one domain's demands can be sacrificed.*

6.3 3G Failures Propagated to 4G System

S6 is a cross-system coupling case found from our experiments. The involved protocols are MM in 3G and EMM in 4G. The usage scenario is to make phone calls in 4G from the LTE phone. In this setting, CSFB is again used. The 4G carrier thus uses its legacy 3G system for the call. During the inter-system changes due to CSFB, location updates are performed in both 3G and 4G. However, such updates may fail. In both OP-I and OP-II, the error message on location area update failure in 3G is propagated to 4G. The 4G user may consequently become out of service, and the operator gains no benefit. Note that, location update is triggered during periodic refresh or CSFB calls, in addition to user mobility. The problem appears to be partly due to improper operational practice, and partly due to the standards that fail to specify the procedure.

Two location updates in 3G are performed when using CSFB for voice calls. The first update is needed after the 4G→3G switching once the call starts. It is initiated by the device. The standards state that this update action can be deferred until the call completes [5]; this helps to reduce the latency when serving the call in 3G. When the call completes, the second location update in 4G is done after the device switches back to 4G. It is done by the network. The update is first processed by MME in 4G, which relays the update request to MSC in 3G. Therefore, based on the standards, two location updates in 3G are activated.

Among the two location updates, one is deemed redundant. It yields no benefit, but incurs penalty. Which specific update does harm depends on the carrier. In OP-I, the first update hurts. The reason is that the delayed update is done once the call terminates. Since the inter-system switching back to 4G is fast, the device-initiated first update is disrupted. This incomplete update status is propagated from 3G to 4G, which sends the device a message with error type "*implicitly detach*". Upon receiving the error, the device enters the "*out-of-service*" state. Note that the 3G system already completes the second update, and the first one is unnecessary. In OP-II, the second update causes damage. The first update is completed first, since it takes more time for the carrier to switch from 3G back to 4G. The success of the first update may trigger MSC in 3G to refuse the second update that is relayed by MME to 3G. It thus replies to 4G MME with an error type "*MSC temporarily not reachable*". A detach request is sent by 4G to the device, and user enters the "*out-of-service*" state.

Problem	S1	S2	S3	S4	S5	S6
Observed	√	×	√	√	√	√
Occurrence Prob.	3.1% (4/129)	0.0% (0/30)	62.1% (64/103)	7.6% (6/79)	77.4% (113/146)	2.6% (5/190)

Table 5: Summary of user-based study on S1-S6.

Operator	Min	Median	Max	90th percentile	Avg
OP-I	1.1s	2.3s	52.6s	13.7s	6.2s
OP-II	14.7s	24.3s	253.9s	34.7s	39.6s

Table 6: Duration in 3G after the CSFB call ends (S3).

Note that both carriers make their decision with plausible excuses. If location update in 3G fails, it does harm the 4G LTE user. The user may miss *incoming* calls. Such incoming calls cannot reach the mobile user if its location update fails. This is why both carriers share and act on the error messages regarding location update failures in 3G and 4G. However, this error-handling process should be confined between 3G MSC and 4G MME inside the network infrastructure. Indeed, they can collaborate to resolve the failures. The error-handling actions should not be directed and exposed to the device. This malpractice can be avoided.

Insight 6: *For the same functions in different networks, they should be coordinated to reduce the conflict. Particularly, the internal failure from one network should not be propagated to another network.*

7. USER STUDY

To assess the real-world impact, we conduct two-week user study with 20 volunteers, including students, faculty members, engineers and technology-unsavvy people. 12 people use 4G-capable phones, while others use 3G-only phones. We observe 190 CSFB calls, 146 CS calls in 3G, 436 inter-system switches (380 switches are caused by 190 CSFB calls), and 30 attaches induced by (re)starting user devices or auto recovery from the *out-of-service* state. Table 5 summarizes the results for six instances S1-S6.

S1 (§5.1): In S1, a user in 3G fails to switch to 4G if its PDP context is deactivated. In our study, we observe 218 4G→3G switches due to CSFB calls (190), user mobility (10) and carrier operations[6] (8). 129 of them are made while mobile data is ON, and 4 S1 events are observed. This results in about 3.1% (4/129) for S1 events in case of 4G→3G switches with enabled mobile data.

S2 (§5.2): S2 results in the attach failure. 30 attaches are observed but none of them fails. It implies that S2 rarely occurs. This can be due to that all are performed in the area with good coverage (the weakest signal strength is -95dBm).

S3 (§5.3): In S3, users do not immediately return to 4G when a CSFB call ends. Among 190 CSFB calls, 103 (39 in OP-I and 64 in OP-II) are made while mobile data is enabled. Table 6 shows the duration in 3G after their CSFB calls end. OP-I users usually switch back to 4G within 3 seconds. It is because OP-I uses "*RRC Connection Release with redirect*," which can be triggered at RRC Non-IDLE state. However, OP-II users get stuck in 3G much longer because OP-II performs "*inter-system cell selection*," which occurs only at RRC IDLE state. We note that all are shorter than that in validation experiments. This is because the duration of getting stuck in 3G depends on the lifetime of ongoing data sessions.

S4 (§6.1): We mainly consider the HOL blocking for 3G CS calls. We check whether there is any location area update done in 1.2 s right after the outgoing call starts, because this update takes at least 1.2 s to complete (§6.1). We observe 79 outgoing calls out of 146 CS calls in 3G. Six (*i.e.*, 7.6%) are affected. In case of longer location area updates (>1.2 s), the ratio is larger.

[6]Note that it may be still triggered by user mobility. However, we cannot justify it since GPS is not always turned on by participants.

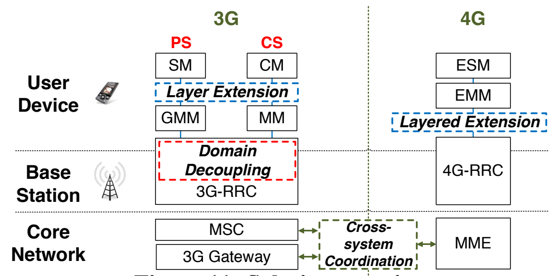


Figure 11: Solution overview.

S5 (§6.2): We examine how often CS calls affect PS data traffic and how much data is affected during a call. It is observed that 77.4% 3G CS calls (113 out of 146) happen while data traffic is ongoing. For these calls, the average duration is 67s, and the average affected data volume is 368KB. Most calls (109/113) affect the data volume less than 550KB, whereas the remaining four calls have impact on more than 4MB data (the largest one is 18.5MB).

S6: (§6.3): In addition to S1, the failure of location update required by CSFB calls make the users fail to switch back to 4G after a CSFB call. It turns out to happen in 5 out of 190 calls (2.6%).

This study with small samples may not accurately quantify the real-world impact and can be further improved with more participants. The result partly confirms that current cellular networks are largely successful. However, it also shows that the found issues do occur in our daily life and affect our real mobile usage. Moreover, though some issues arise with small or negligible probability during normal usage, they may be manipulated and inflated if malicious exploits are launched against cellular networks or users.

8. SOLUTION

We now present our solution, as shown in Figure 11. It has three modules of layer extension, domain decoupling and cross-system coordination. We next elaborate on each component.

Layer Extension. We propose a slim layer with reliable transfer for the out-of-sequence signaling in §5.2 at the EMM, and then parallelize independent operations in §6.1. In the former, the slim layer is inserted between EMM and RRC. Its reliable transfer ensures the end-to-end in-order signal exchange between the phone and MME. To be compatible with the current system, it bridges the interfaces between EMM and RRC and encapsulates the information of reliable transfer function. For the latter, location update should be decoupled from the CS or PS service request for MM and GMM, respectively. Each of MM/GMM maintains two parallel threads. One is for the location update, whereas the other is for remaining functions including the outgoing CS/PS service request. The outgoing CS/PS service request is given higher priority than location update, since the former procedure implicitly does the latter.

Domain Decoupling. Two domains are coupled at the RRC layer. Therefore, we propose a domain decoupling module in RRC. It aims to eliminate the unnecessary interference (*e.g.*, triggered events in §5.3, modulation downgrade in §6.2) from one domain to another. For the triggered events, one domain should not be constrained by another domain. That is, when CSFB is triggered in the CS domain, it should perform 3G→4G switch when the call ends. If the switch condition is satisfied (*e.g.*, 4G is available), the switch will be executed, not blocked by the operations in the PS domain. To this end, the base station adds a CSFB tag to assist the subsequent inter-system switching.

To avoid the modulation downgrade, the 3G RRC can decouple PS and CS services by assigning different channels. Therefore, PS and CS services can be transmitted with different modulation schemes (*e.g.*, 64QAM for PS and 16QAM for CS). To enable the decoupling, we distinguish CS/PS traffic and assign radio resource independently. Both can be satisfied within the current standard and system. First, Radio Link Control (RLC, refer to Figure 1) can exploit the source of traffic (different modules and interfaces used for CS and PS) to differentiate voice and data traffic. Second, the standard allows to assign one device multiple radio channels, each of which can be configured separately.

Cross-system Coordination. The similar functions in different systems should be coordinated because they seek to serve the similar purpose, despite using (slightly) different system-specific approaches. The key is to (1) share the information with each other and (2) collaborate to enforce proper operation. Specifically, 4G EPS bearer context and PDP context are equivalently critical to enable data services. Two systems should enforce the proper transition when the user device switches across 3G and 4G. We recommend that one detach condition should be removed in the standard. It is triggered when the user device without active PDP context switches from 3G to 4G. Instead of detaching itself, the device should immediately activate EPS bearer after inter-system 3G→4G switching. Thus seamless system change can be ensured (§5.1).

In case of failures in one system, the other system should help on recovery if possible. For example, in the second issue (§6.3), the 4G MME should not detach the user device upon the failure of location update in the 3G. Instead, it should recover the devices' location update with the 3G MSC on behalf of the device. In the standard, it is not stipulated that the MME should detach the user device upon the 3G failure. We suggest the operators abolish it.

9. PROTOTYPE AND EVALUATION

We describe the solution prototype and assess its effectiveness.

Prototype of Control Plane. We prototype the control plane functions at three major components, user device, base station, and core network in the cellular network. The user device uses a programmable Android phone. We use two commercial machines (both Lenovo X230) to emulate the base station and the core network. Note that our prototype is based on our own proof-of-concept 3G/4G stacks, since the operational stacks are not accessible.

We implement the modules of connectivity management (CC/SM/ESM) and mobility management (MM/GMM/EMM) at both the user device and the core network. For connectivity management, there are two functions: CS/PS service establishment/release, and the activation/deactivation of PDP context/EPS bearer. The mobility management module provides three functions: attach/detach, location update, and signaling establishment of SM/CM/ESM. We also implement the RRC layer at the device and the base station. Since the transmission at the RRC layer is not reliable, we use UDP to emulate it. We use TCP to forward (relay) RRC payloads between the base station and the core network, since their transmission is assumed to be reliable. All functions are implemented in the application layer.

9.1 Layer Extension

We show that our reliable shim layer in §8 prevents the detach caused by the duplicate or the lost EMM signaling messages. To emulate the lost of EMM messages, the RRC at the base station drops the message according to a given drop rate. For each test, user device does both attach and tracking area update for 100 times. Figure 12 (left) shows that the number of detach varies with the given drop rate with/without our solution. Note that the detach times linearly increase with the drop rate when no solution is used. With our solution, there is no detach while the drop rate increases.

To decouple the location update from the CS service, both the device and core network's MM create two threads to handle them

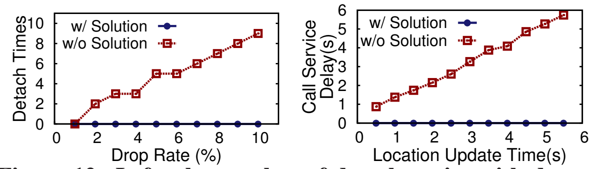


Figure 12: Left: the number of detach varies with drop rate. Right: the call delay call varies with the location update time.

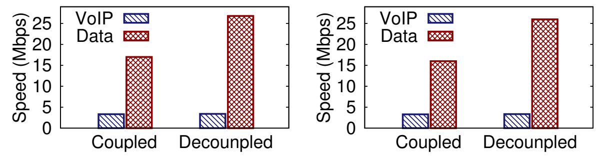


Figure 13: The data speeds vary with/without the coupled data and voice: downlink (Left) and uplink (Right).

concurrently. The location update and the PS service for GMM are also decoupled in the same way. We examine the CS/PS service delay incurred by the location update in MM/GMM. We show only the result of the CS service, and PS service's result is similar. The MM function is configured to do location update every 30 seconds. When the location update is triggered, CM at the user device immediately triggers a call service through MM. Figure 12 (right) shows that the call service delay at MM varies with the processing time of location update. Note that the processing time may vary with the loading of signalling at the core network. Without enabling our solution, the service delay linearly increases with the processing time. However, our solution does not have delay since MM has two threads to deal with location update and call service concurrently.

9.2 Domain Decoupling

We decouple the CS/PS service with two actions. First, we apply different modulations (channels) to CS and PS traffic. Since we have no BS access, we use WiFi Rate Adaptation (RA) module to emulate 16QAM and 64QAM modulation in the CS/PS decoupling case. This can be approximated by using two 48 Mbps and 24 Mbps rates in 802.11a. Note that the overhead could be different between 3G and WiFi, but the result is similar. Figure 13 shows the speed for voice and data in both coupled and decoupled cases. Voice traffic is generated by Skype's VOIP calls. It is observed that the speed of data traffic at the decoupling can be improved by about 1.6 times for both downlink and uplink. In the mean time, the voice can still be carried by a robust modulation. The difference between the speeds of voice and data at the coupling, comes from the voice's small packet size. It incurs more overhead on transmission.

Second, to prevent the CSFB inter-system switching from being blocked in the PS domain, we add a new function into the BS's RRC. It asks the user device to switch its RRC state to a proper state for inter-system switching, once the switching is used to complete the CSFB procedure. It is verified that the user device's CSFB switching is never blocked, by enabling our solution.

9.3 Cross-system Coordination

We prototype two remedies for the cross-system coordination between 3G and 4G. First, the user device always activates the EPS bearer if it does not have active PDP context, after inter-system 3G→4G switching. We test it in the scenario that the user device without PDP context switches from 3G to 4G. The remedy can prevent the device from being detached, so the switch takes only 0.1-0.4s (median is 0.27s). Without the remedy, it takes 0.3-1.3s (median is 0.9s) since the device has to re-attach to 4G network after being detached. This delay may be much larger due to more complicated procedure or the heavy loading at the operator's core network. It is observed as large as 24.7s (§5.1).

In the second remedy, two actions are taken by MME once it receives the failure message of 3G location update for a user device. First, it does not forward this failure message to the device. Second, it triggers the recovery process by updating the device's location to the 3G MSC. It is verified that the MME does not detach the user device upon the failure of location update in the 3G, and further recover it by updating device's location with the MSC later.

10. RELATED WORK

Cellular networking has been an active research area in recent years. Some interesting findings on inappropriate cellular network operations are reported, including the interplay between applications and cellular infrastructure [9,14], TCP over cellular data forwarding [13], mutual interference between data and voice [27], and misbehaviors in cellular functions [21,22,26], to name a few. Our work differs from all such early studies. They focus on packet transmission on the data plane while we work on the control plane. Moreover, they study protocols at the end hosts whereas our entities are at both the devices and the infrastructure elements.

Protocol verification has been investigated for the Internet protocols [12,17,19,24]. New techniques have been recently developed for more complex scenarios. For example, [18] presents SAT-based data-plane debugging; [10] enhances the OpenFlow application debugging with symbolic execution of event handlers; header space analysis is applied for testing complex interactions between various Internet protocols [15]. Our study is orthogonal to these efforts. We focus on protocol verification for cellular networks.

In cellular networks, formal model analysis has been applied to individual protocols. [20] verifies the 2G handover protocol using a generic mobility model with π-calculus. [25] models the authentication protocol and identifies several security loopholes. Our work differs in both the problem and the solution approach. We study protocol interactions and employ two-phrase verification.

11. CONCLUSION

In 3G/4G cellular networks, control-plane protocols are more complex than their counterparts over the Internet. They have to work in more diversified usage settings, e.g., between CS and PS domains, and across 3G and 4G systems. They also support additional functions, including mobility, data and carrier-grade voice, fine control over radio resources. Consequently, inter-protocol signaling is widespread along all three dimensions of cross-layer, cross-domain, and cross-system scenarios.

We show that, some interactions are not well designed, whereas others are not properly operated. The inter-dependent signaling protocols may not take concerted actions. The independent ones are unnecessarily coupled. The incurred damages include both functional incorrectness and performance degradation. The penalty is more pronounced than data-plane faults in data transfer. They may get mobile users stuck in 3G, or deny them 4G access.

Three domain-specific lessons in cellular networks are learnt from our work. First, in the cross-layer case, the well-tested layering rule from the Internet should be honored. If the lower layer does not provide certain functions, the higher layer has to do so, or to be prepared to work without those functions. Coupling inter-layer actions is also not a good practice unless properly justified. Second, in the cross-domain case, signaling design should recognize the inter-domain difference. Treating domains identically seems to reduce design and operational complexity, but makes it overly simplistic and error prone. Third, in the cross-system case, failure messages can be shared and even acted upon between systems. However, it is better not to expose such failure-handling operations outside the system unless absolutely needed.

In the broader scope, research on control-plane protocols in cellular networks warrants more efforts. 3G/4G is a large-scale infrastructure on a par with the wired Internet. There is no competing wireless technology for universal coverage and wide-area mobility support on the horizon. Given such a critical system indispensable to smartphones and tablets, more research is needed. The control-plane research in cellular networks also complements the study on the Internet counterpart. While the Internet seeks to enhance its control plane (*e.g.*, [16]), the cellular system needs to simplify its signaling design. Both can benefit from each other in the process.

12. ACKNOWLEDGMENTS

We greatly appreciate our shepherd, Dr. Ratul Mahajan, and the SIGCOMM reviewers for their valuable feedback. We also thank all participants in the user study. This project at its early stage was partially supported by NSF Grant CSR-1017127.

13. REFERENCES

[1] Speedtest.net - Ookla. http://www.SpeedTest.net.
[2] Voice over LTE. http://www.gsma.com/technicalprojects/volte.
[3] 3GPP. TS25.331: Radio Resource Control (RRC), 2006.
[4] 3GPP. TS23.401: GPRS Enhancements for E-UTRAN Access, 2011.
[5] 3GPP. TS23.272: CSFB in EPS, 2012.
[6] 3GPP. TS24.008: Mobile Radio Interface Layer 3, 2012.
[7] 3GPP. TS36.331: Radio Resource Control (RRC), 2012.
[8] 3GPP. TS24.301: Non-Access-Stratum (NAS) for EPS; , Jun. 2013.
[9] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani. Energy Consumption in Mobile Phones: A Measurement Study and Implications for Network Applications. In *IMC*, 2009.
[10] M. Canini, D. Venzano, P. Peresini, D. Kostic, and J. Rexford. A NICE Way to Test OpenFlow Applications. In *NSDI*, 2012.
[11] H. Holma and A. Toskala. *WCDMA for UMTS - HSPA Evolution and LTE*. Wiley, 2007.
[12] G. J. Holzmann. *Design and Validation of Computer Protocols*. Bell Laboratories, 1991.
[13] J. Huang, F. Qian, Y. Guo, Y. Zhou, Q. Xu, Z. M. Mao, S. Sen, and O. Spatscheck. An in-depth study of LTE: Effect of network protocol and application behavior on performance. In *SIGCOMM'13*.
[14] U. Javed, D. Han, R. Caceres, J. Pang, S. Seshan, and A. Varshavsky. Predicting Handoffs in 3G Networks. *SIGOPS Oper. Syst. Rev.*, 45(3):65–70, Jan. 2012.
[15] P. Kazemian, G. Varghese, and N. McKeown. Header Space Analysis: Static Checking for Networks. In *NSDI*, 2012.
[16] T. Koponen, M. Casado, N. Gude, et al. Onix: A distributed control platform for large-scale production networks. In *OSDI*, 2010.
[17] B. T. Loo, J. M. Hellerstein, I. Stoica, et al. Declarative routing: Extensible routing with declarative queries. In *SIGCOMM'05*.
[18] H. Mai, A. Khurshid, R. Agarwal, et al. Debugging the data plane with anteater. *SIGCOMM Comp. Comm. Rev.*, 41(4):290–301, 2011.
[19] M. Musuvathi and D. R. Engler. Model checking large network protocol implementations. In *NSDI*, 2004.
[20] F. Orava and J. Parrow. An algebraic verification of a mobile network. *Formal Aspects of Computing*, 4(6):497–543, 1992.
[21] C. Peng, C. Li, G. Tu, S. Lu, and L. Zhang. Mobile Data Charging: New Attacks and Countermeasures. In *CCS*, 2012.
[22] C. Peng, G. Tu, C. Li, and S. Lu. Can We Pay for What We Get in 3G Data Access? In *MobiCom*, 2012.
[23] Qualcomm. Circuit-Switched Voice Services over HSPA.
[24] M. A. Smith. Formal Verification of Communication Protocols. In *FORTE*, pages 129–144, 1996.
[25] C. Tang. *Modeling and Analysis of Mobile Telephony Protocols*. PhD thesis, Stevens Institute of Technology, 2013.
[26] G. Tu, C. Peng, C. Li, et al. Accounting for Roaming Users on Mobile Data Access: Issues and Root Causes. In *MobiSys*, 2013.
[27] G. Tu, C. Peng, H. Wang, C. Li, and S. Lu. How Voice Calls Affect Data in Operational LTE Networks. In *MobiCom*, 2013.