How Voice Calls Affect Data in Operational LTE Networks

Guan-Hua Tu, Chunyi Peng, Hongyi Wang, Chi-Yu Li, Songwu Lu
University of California, Los Angeles, CA 90095, USA
{ghtu, chunyip, hywang, lichiyu, slu}@cs.ucla.edu

ABSTRACT

Both voice and data are indispensable services in current cellular networks. In this work, we study the inter-play of voice and data in operational LTE networks. We assess how the popular CSFB-based voice service affects the IP-based data sessions in 4G LTE networks, and visa versa. Our findings reveal that the interference between them is mutual. On one hand, voice calls may incur throughput drop, lost 4G connectivity, and application aborts for data sessions. One the other hand, users may miss incoming voice calls when turning on data access. The fundamental problem is that, signaling and control for circuit-switched voice and packet-switched data have dependency and coupling effect via the LTE phone client. We further propose fixes to the identified issues.

Categories and Subject Descriptors

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

Keywords

Cellular Networks; Mobile Data Services; Voice Call

1. INTRODUCTION

Voice and data are both services indispensable to cellular networks. The IP-based, mobile data access is vital to the surge of smartphones and tablets. In parallel, cellular voice has been the killer application to carriers and users for years. In a legacy 3G network, voice calls are supported via the circuit-switched (CS) path, and data are offered via its packet-switched (PS) route. However, the 4G LTE[1] technology decides to use PS delivery only. This is good news for IP-based mobile data, but poses challenges for voice support. In the absence of CS delivery, LTE carriers have been using a popular solution to voice service, called circuit-switched fallback (CSFB) [1]. CSFB leverages the deployed 3G/2G infrastructure and its CS delivery. It relays LTE voice calls to the legacy 3G/2G networks and enables CS-based voice service. Due to its simplicity and no extra deployment cost, most carriers, including four of the top-five worldwide operators [2, 3, 7, 8] and two major US carriers, have deployed or are planning to deploy CSFB.

In this work, we study the inter-play between voice calls and data services in operational LTE networks. Our objective is to understand how well the CSFB voice works with the PS data over LTE. We are interested in identifying scenarios where they may interfere with each other in both expected and unanticipated manners. We quantify their mutual impact, identify root causes, and design solution fixes.

At a first glance, the above problem seems to be either ill-posed or pretty trivial. In fact, the 4G LTE networks used by data service and the 3G/2G networks used by CSFB voice are indeed independent in operations. Voice and data thus have little mutual dependency. The only merging point is that, the 4G LTE phone has to switch its radio back to 3G/2G networks during a voice call. While it is anticipated that ongoing data may suffer from reduced throughput during the call, there is little beyond this expected performance degradation. However, our study shows this is not true.

Our experiments over two US operational LTE carriers (called as OP-I and OP-II) have yielded four findings, one expected and three unanticipated. First, we indeed observe throughput slump for data sessions up to 83.4% when the 4G LTE user falls back to 3G for voice calls. This drop is caused by the speed gap between LTE and 3G, but also incurred by data suspension and losses during CSFB-triggered handoffs between 4G and 3G. The good news is that this degradation occurs mainly during the voice call for OP-I ; However, for OP-II, the degradation may last even after the call ends (see Finding 2). Second, we discover that 4G LTE users may lose 4G connectivity due to voice calls. They will not return to 4G afterwards. The lost connectivity lasted more than 10 hours and showed no sign of limit. The issue occurs when certain background data traffic is running in some voice call scenarios. In particular, it happens when the voice call fails to be established (for OP-I), or no matter if the call is established (for OP-II). We identify that it is caused by the state machine "loophole" that 3G is unable to switch back to 4G under certain scenario. Third, data applications may abort (about 2-5% on average and 15% in the worst case in our tests) when a voice call ends. The network may implicitly detach the user, despite ongoing data sessions, when migrating the user back to 4G after CSFB calls. Consequently, the state or signaling triggered by CS voice also affects PS data service. Last, we discover that PS data may also affect CS voice. CS calls may not be available when the mobile turns on its PS service. The network state can then be changed by PS, thus leading to transient unavailability of CS. Table 1 summarizes all these four findings.

The above findings confirm that, the interference between voice and data in CSFB-capable LTE is mutual. Although these experi-

[1]In this paper, we will use 4G and LTE interchangeably.

MobiCom'13, September 30–October 4, Miami, FL, USA.

http://dx.doi.org/10.1145/2500423.2500429.

Finding	Operators	Detail	Root Cause	Section
Throughput slump	OP-I, OP-II	Data throughput decreases (up to 83.4% observed); OP-I: only during the call, OP-II: during and after the call	Handoffs triggered by CSFB and speed gap between 3G and 4G	Section 4
Losing 4G connectivity	OP-I, OP-II	Never returns to 4G after the CSFB call under certain data traffic; OP-I: when the call fails to be established, OP-II: any CSFB call	State machine "loophole" in 3G→4G transition	Section 5
Application aborts	OP-I, OP-II	Application aborts occasionally (3.4% for OP-I, 5.7% for OP-II) after the call;	Network state changed by CS-domain operation (here, network detach caused by CSFB voice calls)	Section 6
Missing incoming call	OP-I, OP-II	Misses all incoming calls temporally (for several seconds) while enabling the PS service	Network state changed by PS-domain operations	Section 7

Table 1: Finding summary.

mental cases do not necessarily represent the common usage scenarios, they do showcase worst, yet possible settings. It indeed reveals complicated dependency and coupling effects between voice and data. These effects are induced by the fundamental design of CSFB, as well as its implementation loopholes. We further devise solutions that coordinate with the LTE phone to fix these issues.

The rest of the paper is organized as follows. Section 2 introduces the 4G/3G architecture and voice support via CSFB. Section 3 describes our study methodology and the addressed issues. Sections 4, 5, 6 and 7 present each individual finding and explore its root causes. Section 8 proposes our solution fix. Section 9 discusses alternatives to CSFB. Section 10 compares with the related work, and Section 11 concludes this paper.

2. BACKGROUND

We introduce the 4G LTE architecture and its legacy 3G network. We then describe how voice and data services are provided.

Cellular network: Figure 1 illustrates the LTE architecture, as well as its legacy 3G network, i.e., UMTS (Universal Mobile Telecommunication System). The LTE network offers PS data service. It consists of core network, radio access network (RAN) and user equipments (UEs, i.e., mobile devices). Its RAN uses eNodeB (LTE base station) to offer radio access to UEs. Its network core is IP-based, consisting of MME (Mobility Management Entity) to handle user mobility (e.g., location update or paging UEs), and 4G gateways that route packets between the Internet and the 4G RAN.

In contrast, 3G network supports both PS and CS to offer data and voice, respectively. Its RAN uses RNS (Radio Network System) to provide radio access. Its core network has several main elements: (1) GMSC/MSC (Gateway Mobile Switch Center), which pages and establishes CS services (e.g., voice calls) with mobile users; (2) HLR/VLR (Home/Visitor Location Register), which stores user information (e.g., location updates); (3) 3G gateways, which route PS packets between the internet and the RAN. There exist 3G technology variants, such as 3G HSPA (High Speed Packet Access) that offers data rate up to 14.4–42 Mbps. Note that 3G architecture is also applicable to 2G; We only describe 3G since 2G is seldom observed in our study.

Voice calls (and data) for 4G LTE users: Since the LTE network uses PS only, it is unable to use CS to support voice, which traditionally requires guaranteed service quality. Instead, LTE uses CSFB that supports voice calls in the legacy 3G CS network. LTE also claims VoLTE (voice-over-LTE) as its ultimate voice solution [9], to be discussed in Section 9.

In CSFB, when a 4G user is called, the incoming call is routed to the GMSC in 3G networks. GMSC then queries HLR/VLR to learn which MSC the 4G user is located at, and forwards this call to the serving MSC. The MSC subsequently pages the user device through MME. Once the UE is found, MME migrates it from 4G

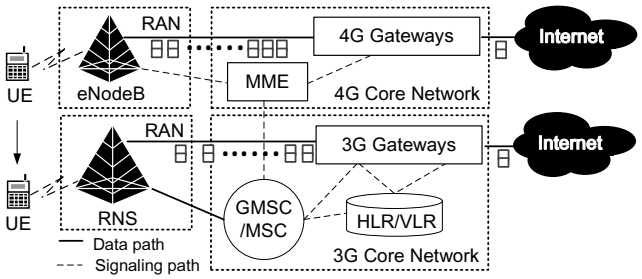


Figure 1: 4G/3G network architecture and CSFB.

LTE networks to 3G networks via triggering inter-system handoff (i.e., from 4G to 3G). After the UE successfully connects to 3G RANs, the MSC establishes the voice call with UE. In the meantime, ongoing data sessions are also transferred to 3G networks together with voice. The outgoing call performs similarly except it sends MME a request to switch to 3G networks.

3. STUDYING CSFB IN OPERATIONAL LTE NETWORKS

In this section, we describe our experimental methodology and identify the key problem aspects to be studied.

3.1 Experimental Methodology

We conduct experiments in two major US LTE operators, denoted as OP-I and OP-II, for privacy concerns. They together serve more than 138M mobile subscribers and cover almost 50% US market share [13]. We use six phone models of LG Optimus G, Samsung Galaxy S3, S4 and Stratosphere, HTC One, and Apple iPhone5, running two mobile operating systems: Android and iOS. For OP-II, we use Galaxy S4 and iPhone5 only. They run popular applications (e.g., YouTube) or conduct data sessions with our deployed servers, including Apache Web server, FTP and TCP/UDP servers. For further performance analysis, our deployed TCP/UDP server adds a sequence number in each data packet to/from the UE. We primarily collect and analyze traces from Android phones, and Apple iPhone5 is used for verification experiments.

In each experiment, we collect five traces if available: (1) *Wireshark:* We use the Wireshark for packet capture traces on mobile devices and our deployed servers. (2) *TcpParms:* We use *getsockopt*, a socket API to periodically log TCP parameters, such as retransmission timeout or congestion window, on both our TCP server and mobile devices (root is required) . (3) *UdpSeq:* To verify whether out-of-order delivery is observed by CSFB-induced inter-system handoffs, we log the sequence number carried in the received UDP datagram and timestamps on our deployed UDP servers and mobile devices. (4) *NetworkStatus:*

Mobile devices also record network status information given by Android `PhoneStateListener` class. The *NetworkTrace* periodically collects phone status information including *timestamp*, *operator*, *network type*, *cell identifier*, *RSSI* (Signal Strength) and *IP address*. The record interval is 100ms. (5) *CallEvents:* Mobile devices also log all incoming-call events on phones via `PhoneStateListener` and outgoing-call events, e.g., ringing, and current timestamp.

3.2 Issues to Study

In operational LTE networks, data service is offered via the IP-based, PS service, while the voice service is provided through mechanisms such as CSFB. Since CSFB is probably the most popular mechanism in practice to support voice in LTE networks, we focus on it in this study. We discuss other alternatives of VoLTE and SVLTE in Section 9.

Conventional wisdom states that such data and voice will not *interfere* each other, or at least not to the degree beyond expectation. Anyway, data is going through the 4G LTE infrastructure, while voice is going through the separate 3G/2G networks. However, our study shows that this is not the case. We carry out our research along both directions: (I) How does CSFB voice affect the ongoing data service in LTE networks? and (II) How does the data session in LTE networks affect the voice service? While the results for (II) are presented in Section 7, the details for (I) need more elaboration and are given in Sections 4 - 6. As data service becomes increasingly important for mobile devices, it deserves more attention. In particular, we cover three aspects, expected, and unexpected, even certain worst-case scenario, regarding how voice affects data in the context of LTE networks:

1. How much is the performance degradation when voice calls occur? This is the somewhat expected case for performance penalty. The data session falls back to 3G/2G networks during a CS voice call and then returns to 4G data networks while the call ends. We seek to understand how TCP and UDP transport protocols react to such scenarios, as well as worse-than-expected instances (Section 4).
2. Can the data session go wrong when call completes or is never established? If it indeed occurs, it will be unanticipated exceptions for CSFB. We seek to show certain extreme cases of losing LTE connectivity and getting stuck in 3G even when voice calls complete or never start and explore their root causes (Section 5).
3. Can voice calls incur other negative performance impact beyond throughput degradation? In particular, we will illustrate cases of application abort when voice calls are underway and identify their root causes (Section 6).
4. Can the PS data also affect the CS voice call under certain conditions? If it is indeed observed, it shows that both data and voice have mutual interference on each other's operations (Section 7). We also explore its root cause.

Table 1 summarizes our findings over two US carriers on the above four issues. We elaborate them in Sections 4 to 7.

4. THROUGHPUT SLUMP

In this section, we first examine how data performance is affected by voice calls using CSFB in the *normal* case. The user might experience throughput slump during voice calls due to the handoff from 4G to 3G. This observation matches our expectation and recent reports [10]. We elaborate on what happens to TCP/UDP based data sessions and study the impact of regular CSFB calls. We finally report *worse-than-expected* findings: performance degradation under multiple handoffs (OP-I) or even *after* the voice call (OP-II).

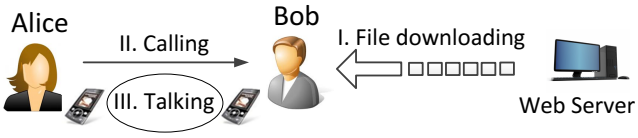


Figure 2: Alice calls Bob while he is downloading a file.

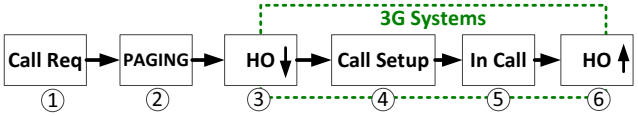


Figure 3: CSFB event flow for an incoming call.

4.1 An Illustrative Example

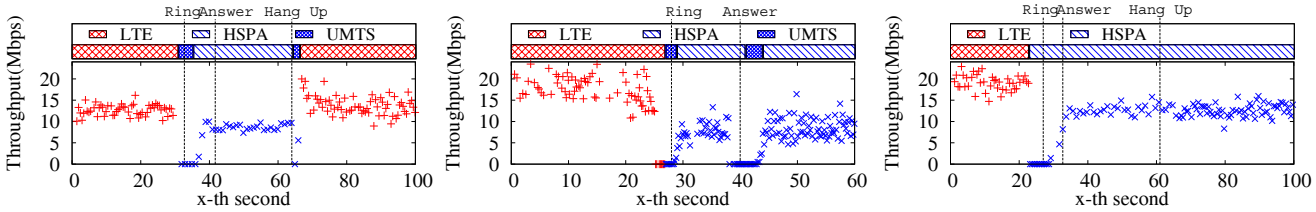
We use an example to illustrate the normal case performance. Bob is downloading a file to his Samsung Galaxy S3 via the high-speed 4G LTE network. Everything goes well until he receives a call from Alice. The call lasts about 22 seconds. The procedure is illustrated in Figure 2. Figure 4(a) records the data throughput, network type, and call events observed at Bob's phone using OP-I carrier. At the beginning (up to 29.8th second before the call), Bob enjoys high-speed data access up to 14 Mbps; During the voice call (during the interval [42s, 64s]), the throughput drops from 14 Mbps to 9 Mbps; When the call ends (about 64th second), the throughput increases to 14 Mbps in about 2 seconds. That is, the data throughput decreases by 35.7% (i.e., $(14-9)/14$) *during* the call.

Cause: The observed throughput slump is caused by CSFB. Figure 3 shows the CSFB event flow for an incoming call. We make four observations. First, when answering the incoming call, a handoff procedure from 4G to 3G is triggered. This inter-system handoff takes place (Step 3) even *before* the call is fully established (Step 4). Figure 4(a) shows that, Bob's phone call starts ringing around 33th second and is answered at 42th second. In contrast, the first handoff (LTE to 3G UMTS) completes at 31st second, earlier before the events of ringtone and call answering. Interestingly, at 35.4th second right after the first handoff, the network performs a second handoff (UMTS to 3G HSPA), which upgrades to higher-speed 3G HSPA networks (14.4 - 42Mbps theoretically). Second, the call proceeds during [42s, 64s] until the call hangs up. The phone stays in the 3G HSPA network during this period. Third, once the call completes, the phone switches back to 4G after two handoffs (HSPA to UMTS, followed by UMTS to 4G) at 65th and 66.4th seconds. Note that, the 4G CSFB standard [1] does not require that users be immediately switched back to 4G after the call ends, or how many handoffs be triggered to switch to 4G. The sixth step is OP-I's implementation choice. We will see OP-II's behaviors later. Last, we observe two data transmission suspensions (i.e., rate is 0 Mbps): 6.4 seconds during [29.8s, 36.2s] and 1.5 seconds during [64.3s, 65.8s]. Both periods are accompanied with handoffs. These handoffs lead to data transmission suspension [1]. Once the handoff is completed, the data transmission resumes.

We next address two issues: (1) How does TCP/UDP react to the above case? (2) Is there any worse-than-expected result, except the performance degradation caused by staying in 3G during the call? In particular, is there any difference between the handoff triggered by CSFB calls and the traditional, mobility-induced handoff?

4.2 TCP/UDP under Normal Voice Calls

TCP: In the above example, TCP data transmission is suspended during [29.8s, 36.2s] and [64.3s, 65.8s]. Figure 5 plots TCP logs



(a) OP-I (b) OP-I, more handoffs (c) OP-II

Figure 4: Logs of data throughput (4G:+, 3G:×), network type (LTE, HSPA, UMTS) and call event (marked by black dashed lines) observed at Bob's phone in normal case of answering Alice's call. (a) OP-I: one 4G→3G handoff triggered; (b): OP-I: multiple handoffs triggered; (c): OP-II: no handoff back to 4G when the call ends.

in [29s, 38s] at the TCP server in the example of Figure 4(a). Note that the server clock is slightly out of sync (about 0.2s to 0.3s) from the mobile's trace. We make three observations on the TCP trace. First, no packet delivery during handoffs results in multiple TCP timeouts. Around the 29.7th second, no ACKs are received for the packet with sequence number 44636389. Accordingly, the server retransmits it four times (at 29.7s, 30.6s, 32.3s, 35.8s, respectively). Second, large RTO may impede fast TCP recovery. The retransmission timeout (RTO) gradually doubles, here, 0.436s, 0.872s, 1.744s, 3.488s, 6.976s during [29.1s, 35.6s]. Large RTO values imply that TCP responds slowly once the network connection resumes. In this case, the fourth retransmission succeeds (another packet sent at 35.9s) and the suspension lasts around 6 seconds. Third, the TCP congestion window is about 244 MSS during [29s, 36s]. It does not reduce immediately upon retransmission timeout, thus different from the TCP specification (RFC 5681). The congestion window update is deferred when data transmission resumes. We believe this is a TCP implementation variant in Linux.

UDP: We observe behaviors similar to TCP, except that the suspension time for UDP is shorter. Since UDP does not have congestion and flow control mechanisms, its transmission resumes immediately after the PS service is available. In contrast, TCP RTO may not expire yet though the PS service resumes. We conduct experiments to test this hypothesis. Before a voice call comes, we start a 100 Kbps UDP downlink session and a TCP downlink flow on our 4G phone. As expected, average data suspension durations for UDP and TCP are 5.4 and 6.4 seconds, respectively. It takes longer for TCP to resume its transmission. We further observe out-of-order data delivery upon 4G→3G and 3G→4G handoffs.

4.3 Worse Than Expected

As expected, data performance degrades during the voice calls due to the speed gap[2] between 4G and 3G and data suspension during handoffs triggered by CSFB. Next, we are curious about whether any worse-than-expected results happen. We uncover two cases of further performance degradation: (1) due to more handoffs (OP-I), and (2) even after the call (OP-II).

More handoffs (OP-I): Handoffs are critical to data performance. Upon handoff, data transmission suspends, thus incurring TCP/UDP throughput decrease. Each CSFB call triggers two network switches: 4G → 3G upon call arrival and 3G → 4G after the call ends. In OP-I (Figure 4(a)), one 4G → 3G switch is enabled by two handoffs of LTE→ UMTS (before the phone rings) and UMTS→ HSPA (before the call is answered).

We next examine whether there is any difference between the handoff triggered by CSFB voice calls and the conventional handoff induced by mobility. Our study shows that the difference indeed

[2]More measurements can be found in Section 5.4.

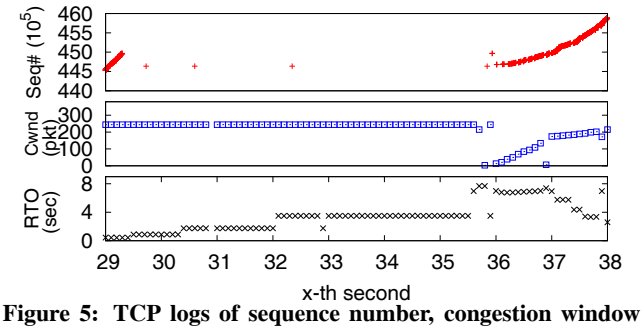


Figure 5: TCP logs of sequence number, congestion window and retransmission timeout observed at Bob's TCP Server in the example of Figure 4(a).

exists; More handoffs may be triggered by call-related events. We run a 100 Kbps UDP session while answering an incoming voice call. We repeat this experiment for 367 runs. In 218 runs, the handoff results are the same as the above example. However, in the remaining 149 runs, two additional handoffs (HSPA→ UMTS and UMTS→HSPA) are triggered by the call-answering operation, as shown in Figure 4(b). Consider the speed of HSPA and UMTS (up to 14.4-42 Mbps and 2 Mbps for HSPA and UMTS, respectively). The mobile user suffers from another performance drop at around 40th second. Note that, the additional HSPA↔UMTS handoffs differ from the mobility-induced one. It is triggered by a call-answering event while the phone remains at the *same* location, performing data and voice services. In contrast, the traditional handoff to 3G UMTS typically occurs upon mobility (i.e., users move out of a HSPA cell).

No handoff back to 4G (OP-II): We observe similar performance drop during the call in both carriers. However, after the call ends, data throughput still remains low for OP-II, different from the throughput increase for OP-I. Figure 4(c) plots the mobile trace at Bob's phone using OP-II. In this example, Bob experiences a throughput slump from 19Mbps to 12.7Mbps during the call [31s, 61s]. However, the throughput still remains around 12.7Mbps after the call. We observe similar behaviors with different phone models (e.g., Galaxy S4 and iPhone5): the handoff occurs before the phone rings and the throughput remains similar after the call ends. Undoubtedly, it adversely imposes larger impact on data throughput. The mobile loses its 4G connectivity even after the voice call. This occurs because no handoff is invoked immediately after the call ends. We will explore its root cause in Section 5.

We further explore why these handoffs are invoked and whether they can be eliminated for performance improvement. Unfortunately, the inter-system handoff (4G→3G) is mandatory to support CSFB in order to access the 2G/3G circuit-switched service for voice calls. The additional handoff (HSPA ↔ UMTS) is the OP-

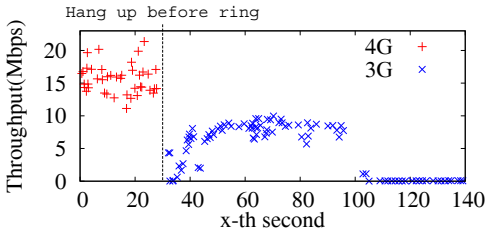


Figure 6: Data throughput observed at Bob's phone if Alice immediately hangs (the outgoing call) up in OP-I carrier.

I's implementation choice; it is not required in OP-II. The handoff back to 4G is also part of the operator's design choice. The CSFB standard never specifies when to switch back to 4G networks. In practice, OP-I decides to perform this handoff immediately, while OP-II does not.

5. LOSING 4G CONNECTIVITY

We observe that LTE users permanently lose 4G connectivity due to CS voice calls under certain conditions. In an extreme test scenario, the mobile phone is stuck in 3G networks longer than 10 hours and shows no sign of exiting. It remains in 3G networks even when the user drives on the route with stronger 4G signal. The condition of losing 4G connectivity varies among two operators. Compared with OP-II, OP-I has more complex settings.

5.1 OP-I: When the Call Fails to Establish

Our test scenario can be illustrated via the Alice-Bob example. However, after Alice calls Bob, she *immediately hangs up* because she realizes that it is too late to call Bob at 10PM. For Bob, his phone never rings and the call is never established. Assume that Bob has been downloading a file before the call. Contrary to our expectation, we discover that Bob gets stuck in the 3G network for a long time (or even unlimited duration). Figure 6 plots the data throughput measured at Bob's phone. It shows that Bob never returns to the 4G network even after the download halts at 100th seconds. The same is observed on all phone models using OP-I. Note that only some background data service keeps on running. Throughout this process, Bob is not even aware of what happens!

We now explore the root cause to lose 4G connectivity. It turns out that a loophole (in fact, a loop) in Radio Resource Control (RRC) state transition forces the 4G user to remain in 3G. RRC is the function that regulates the connection establishment and release between the UE and the core network.

Figure 7(a) plots the simplified RRC state transition in 3G/4G standards [1]. We do not consider 2G here since it is not observed in our experiments. We make two observations. First, the switch between 3G and 4G networks is enabled via the handover procedure (that occurs between 3G `FACH/DCH`[3] and 4G `CONNECTED` states) or the cell reselection procedure [1] (that is invoked between 3G `IDLE` and 4G `IDLE` states). Second, within 3G or 4G, the state transition (e.g., 3G `FACH/DCH` ↔`IDLE` or 4G `CONNECTED`↔`IDLE`) is determined by the connection establishment/release. For example, a RRC connection shall be established before the PS/CS service is used, or be released when the CS/PS service is in no use or remains idle for a long time.

The above RRC state machine brings an inherent risk of getting stuck in one network (e.g., 3G). In case the loop between 3G `FACH/DCH` and 3G `IDLE` is formed, the mobile user will be unable to escape from 3G. Unfortunately, our experiments confirm that such a loop indeed exists under certain conditions in OP-I. In particular, for an unestablished call, the RRC enters into the 3G loop with some ongoing data services.

Unestablished Call State: In the normal case, the mobile user moves back to the 4G network quickly (in about 2–4 seconds) after the call ends. However, the time prolongs if the call is not established. It occurs in two scenarios. One is that the 4G user is called but the caller hangs up immediately (usually within 4–6 seconds). The other is that the user makes an outgoing call and immediately hangs up after the handoff to the 3G network is done. In both cases, the mobile phone falls back to the 3G network though no call has been established.

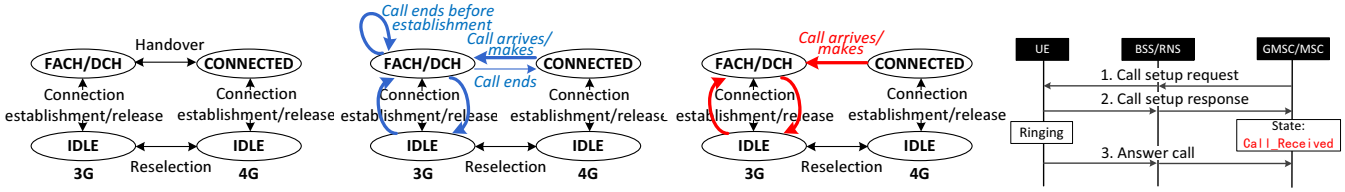
The unestablished call state does make it longer move back to the 4G network. Figure 7(d) shows the call setup procedure (Step 4 of Figure 3) for the incoming call case. When the MSC sends a call-setup request to a 4G phone, it waits for the response from the phone in order to update its state as `Call_Received`. However, when the call is canceled before entering the `Call_Received` state (in the above two scenarios), the MSC will *not* update its call state. The user thus stays longer in the 3G network. This implies that the call state plays a critical role in the handoff operation. The unestablished call changes the trigger condition for handoffs, thus taking longer time to go back to 4G.

The duration to stay in 3G is largely independent of locations and phone models. We test with all phone models at four locations with different base stations. Each test repeats 20 runs. In the absence of data service, the duration to remain in 3G ranges from 7 to 8 seconds, varying slightly with locations. With background data traffic, we observe similar results independent of locations. We figure out that the duration is determined by other factors to be discussed in Section 5.3.

Data Services in Parallel: We discover that, the phone can stay in the 3G network for an extended period of time if some data service is ongoing in parallel. We run a large number of tests to study when and under what conditions the switch (back to 4G) takes place. We run the unestablished call experiments with a constant-rate UDP *uplink* session to our deployed server on Samsung Galaxy S3 and LG Optimus G. We vary the inter-packet spacing (i.e., packet interval) from 1 to 24 seconds, using 1B and 1KB UDP payload sizes. Each test repeats 20 runs. We observe similar results for both phone models, and only describe the results on Samsung Galaxy S3. Figure 8 plots the duration in 3G since the phone moves to 3G. The upper and lower lines mark the maximal and minimal durations in 3G. The 180-second duration implies that the phone never returns to 4G in 3 minutes. It clearly shows that the phone might get stuck in 3G under certain conditions. The condition specifics will be elaborated in Section 5.3.

In summary, we infer the RRC state transition machine for OP-I in Figure 7(b). The 4G→3G handoff is triggered when the call arrives (or is initiated), and the 3G→4G handoff occurs when the call ends after its establishment. However, if the call is not established (i.e., hanging up too early), the 3G→4G handoff will not be invoked. Note that, no handoff for the unestablished call is not designed without rationale. If the call is not successfully established, the caller would probably redial shortly. For a call terminated in the unestablished call state, immediate handoff from 3G to 4G could trigger more handoffs. Consequently, the UE still stays at the 3G `FACH/DCH` state. In this case, the cell reselection procedure turns out to be the only way back to 4G. Note that the cell reselection is only triggered in the 3G `IDLE` state. Our study demonstrates that

[3]FACH and DCH are two RRC states that offer RRC connections for data delivery. FACH offers a RRC connection at lower speed with low power consumption, whereas DCH offers it at full speed with higher power consumption [1, 27].



(a) 3G/4G standard [1] (b) CSFB for OP-I (c) CSFB for OP-II (d) Call setup procedure

Figure 7: Simplified RRC state transition machine and call setup procedure.

under certain data operations (the details are in Section 5.3), the UE may not enter the 3G IDLE state, or switch back to 3G FACH/DCH before triggering the cell reselection. Consequently, a RRC loop (marked by bold blue lines in Figure 7(b)) is created when the call is not established under certain background data traffic.

5.2 OP-II: Once the Call Attempt is Made

Losing 4G connectivity becomes easier in OP-II than in OP-I. The user is prone to losing 4G connectivity, no matter whether the call is established or not. Figure 4(c) shows an example of losing 4G connectivity after the call completes. We test various call operations (answering/rejecting an incoming call, unestablished incoming call, or making an outgoing call) to examine its dependency on the call event. We find out that, once the 4G→3G handoff is triggered under any call attempt, the UE gets stuck in 3G as long as some background data traffic is present.

Similar to the OP-I test, We run constant-rate UDP *uplink* session tests with different packet sizes and intervals. The only difference is that the call is established and completes in this experiment. Figure 9 plots the duration being stuck in 3G with packet intervals for 1B/1KB packets. We observe that, the rule in OP-II is much simpler. When the packet interval is smaller than 9s (1B packet) or 13s (1KB packet), the 4G user is unable to return to the LTE network.

Consequently, we deduce the RRC state transition for OP-II in Figure 7(c). Different from OP-I, no handoff path (back to 4G) exists when the call ends. The switch from 3G to 4G is thus invoked only through the cell reselection. Similarly, when the RRC loop in 3G FACH/DCH and 3G IDLE (marked by red bold lines) is formed, the UE loses its 4G connectivity. The conditions to form the 3G RRC loop will be discussed in Section 5.3.

5.3 RRC Loop Under Data Services

We now examine when or under what data services the 3G RRC loop is formed. We mainly address the OP-I case that is more complicated, and then describe the OP-II case.

5.3.1 OP-I Case

Figure 8 plots the 3G duration under various packet intervals for OP-I. We make four observations. First, when the packet interval is smaller than certain threshold (10 seconds for 1B packets and 15 seconds for 1KB packets), the phone never returns to 4G. Second, when the packet interval is larger than another threshold (15 seconds for 1B packets and 20 seconds for 1KB packets), the phone definitely returns to 4G. Third, for both packet sizes, there exist two interesting transition intervals. For 1B packets, these two intervals are 10 seconds and 15 seconds (with larger duration variance). The phone is likely to return to 4G with these packet intervals. For packet intervals in between (i.e., 11–14s), the phone never returns to 4G. For 1KB packets, the pattern is similar but the two transition intervals are 15s and 20s. Finally, compared with no data transmission, it stays longer (about 40 seconds) in 3G even when it can return to the 4G network.

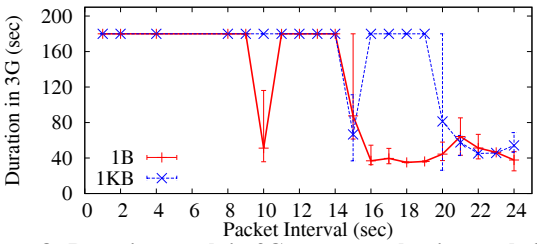


Figure 8: Duration stuck in 3G versus packet intervals for two 1B/1KB packets in case of an unestablished call via OP-I.

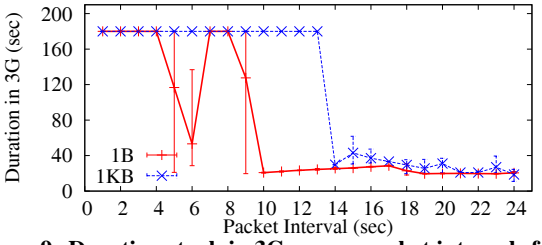
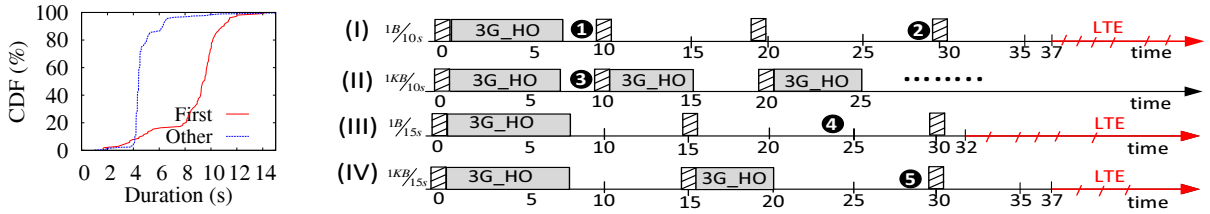


Figure 9: Duration stuck in 3G versus packet intervals for two 1B/1KB packets in case of a complete call via OP-II.

At the first glance, these findings are not anticipated, particularly the inconsistent performance with packet intervals in the transition zone. Three questions need to be answered: (1) Why does the switch remain *sensitive* to several packet intervals and yield the non-monotonic pattern for all packet intervals? (2) What occurs when the packet interval is set as 10s, 15s, and 20s? (3) How is it related to packet sizes? We examine event traces and finally derive the trigger conditions for the 3G RRC loop. We summarize these rules for $3G \rightarrow 4G$ switch in OP-I in Table 2, and then use our trace analysis to explain what happens and how each rule is applied. In summary, the above observations reveal how such mechanisms interact with each other.

These rules exhibit both standard specifications and carrier-specific operations. They also correspond to the state machine derived in Section 5.1. Note that the UE can be switched from 3G RRC IDLE to 4G RRC IDLE only via the cell re-selection procedure. Rule 1 states that this cell reselection is triggered by a timer $T_{3G \rightarrow 4G}$, which is set to 5s according to our measurements. Note that the timer $T_{3G \rightarrow 4G}$ is not specified by the standards, but chosen by the operator's implementation.

Rule 2 regulates the traditional 3G RRC transition between DCH/FACH and IDLE. The transition is controlled by another timer T_{idle}. When this timer expires, RRC jumps from DCH/FACH to IDLE; Upon packet delivery, it immediately switches to DCH/FACH and resets T_{idle}. In our experiments, we find that $T_{idle} = 10$ seconds and is operator specific, consistent with prior studies [27].



(a) Intra-3G handoff time (b) Example traces

Figure 10: Illustration of event traces for data flows with various packet intervals and packet sizes.

Rule 1	The phone immediately performs the switch back to 4G when the timer $T_{3G\to 4G}$ times out; $T_{3G\to 4G}$ is only started when the UE enters into the 3G `IDLE` state;
Rule 2	The RRC state switches from `DCH/FACH` to `IDLE` when the timer T_{idle} times out; It switches to `FACH/DCH` immediately upon any data delivery;
Rule 3	$T_{3G\to 4G}$ is reset once an intra-3G[a] handoff occurs.
Rule 4	The intra-3G handoff occurs when a data transmission request occurs in either condition: (1) the UE is in 3G `IDLE` state, (2) the packet size is larger than a threshold (210-220B in our measurement) or for the first packet.

Table 2: Rules for $3G \to 4G$ switch upon an unestablished call (i.e., the call state is not `Call_Received`) for OP-I.

[a]In this work, we define the intra-3G handoff as handover events within different types of 3G networks, i.e., among UMTS/HSDPA/HSPA/HSPA+ [1].

Rule 3 implies that the users will not go back to 4G LTE when it is at the `FACH/DCH` state, which is specified in [1]. This is because the user triggers an intra-3G handoff to perform data transmission (Rule 4), its RRC state is hence changed from `IDLE` to `FACH/DCH`. The timer $T_{3G\to 4G}$ should be reset since the user leaves the `IDLE` state.

In Rule 4, we observe that intra-3G handoffs (i.e., HSPA↔UMTS) might happen. The operator usually switches the mobile device to proper radio access technologies (e.g., UMTS or HSPA) based on its transmission rate and data volume. The first condition in Rule 4 is obtained from our traces. It is not in the standard specifications; we believe it is also an operator-dependent choice.

Our measurements also indicate that, an intra-3G handoff typically takes about 5 seconds, but 8–10 seconds for the first time. Figure 10(a) plots the CDF for the intra-3G handover duration. To derive the packet size threshold used in Rule 4, we run experiments using variable-sized payloads at 8-second intervals (i.e., it remains in 3G). It turns out that the payload threshold is 210–220B. These parameter settings are also operator specific.

We briefly illustrate how these rules are applied so that the 3G duration varies with packet intervals as shown in Figure 8. In all our experiments, the first packet is immediately sent out once the phone switches to 3G, shown by those packet boxes at time 0 in Figure 10(b). We look at two easy-to-understand cases, while more cases can be found in Appendix A. In the first case, when the interval is smaller than T_{idle}, the phone never returns to 4G. It has no chance to enter the 3G `IDLE` state and trigger the timer $T_{3G\to 4G}$ back to 4G. In the second case, when the packet interval is larger than 20 seconds (i.e., $T_{idle}+T_{3G\to 4G}+T_{3G-HO} \approx 10+5+5 = 20$), the phone can always return to 4G. No matter whether an intra-3G handoff is triggered or not, the packet interval is large enough to enter the `IDLE` state and trigger the $3G \to 4G$ timeout. This also explains why the transition interval for 1B is 5 seconds smaller than that for 1KB. The decisive factor for the 1B/1KB discrepancy is whether an intra-3G handoff is triggered. By Rule 4, 1KB packet delivery always triggers such an intra-3G handoff as long as it is still in 3G, whereas 1B packet does so only when the RRC state is 3G `IDLE`. The intra-3G handoff takes about 5 seconds. In more cases (all four cases of Figure 10(b) and other combinations of packet intervals and sizes) in Appendix A, we confirm that the duration in 3G is mainly determined by how the RRC state and the two timers of $T_{3G\to 4G}$ and T_{idle} evolve under these rules.

5.3.2 *OP-II Case*

We next derive the trigger conditions for the RRC loop for OP-II. Our trace analysis shows that OP-II follows the above four rules but Rule 4 and the parameters vary slightly. Figure 9 shows that the transition intervals for 1B and 1KB packets are around 9s and 13s, respectively. Our traces indicate that the difference between 1B and 1KB packets lies in whether intra-3G handoffs are incurred; Intra-3G handoffs (HSPA→HSPA+) occur for 1KB packets, but not for 1B packets (except for the first packet, the same as OP-I). Moreover, we test with various packet sizes and find out that the occurrence of intra-3G handoffs only depends on the packet size (here, the threshold is about 940–950B). This slightly differs from Rule 4 for OP-I. In OP-II, Rule 4 works under the second condition (an intra-3G handoff occurs when the packet size is larger than 940-950B or for the first packet). Our measurements show that an intra-3G handoff takes about 4–5 seconds (but 8-12 seconds for the first time), similar to Figure 10(a). We infer that the 3G→4G switch is also controlled by two timers, $T_{3G\to 4G}$ for the cell reselection and T_{idle} for the state transition from 3G `FACH/DCH` to `IDLE`. Based on our measurements, we infer that $T_{3G\to 4G} \approx 6$s, $T_{idle} \approx 3$s. The derivation is similar to that for OP-I, and is omitted due to space limit. Note that, these parameters are operator-specific choices.

For both cases of OP-I and OP-II, some may argue that the loss of 4G connectivity is an operator-specific implementation issue. We admit that the operator's choice does matter. Indeed, the standards do not stipulate under what conditions the handoff/switch should be initiated, though they do specify such handoff/switch mechanisms for CSFB [1]. They leave the flexibility to the carriers. For example, the operator can decide whether a handoff back to 4G is triggered immediately after the call ends or not. However, our study shows that, the loss of 4G connectivity is caused by the flaw (i.e., the state loop) between CSFB and the RRC finite state machine. The interplay of functions and states used in both CS and PS domains results in unanticipated effect. The two timers create the 3G RRC loop so that the phone never returns to 4G.

5.4 Performance Impact

We examine the impact of being stuck in the 3G network from three aspects: duration in 3G, mobility, and throughput gap.

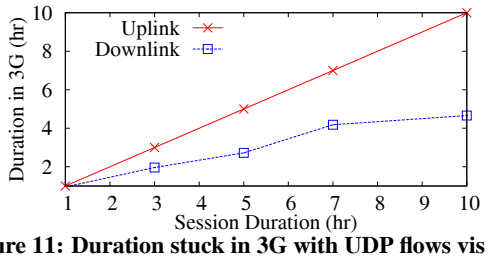


Figure 11: Duration stuck in 3G with UDP flows vis OP-I.

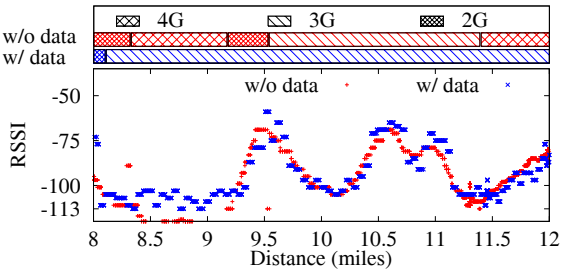


Figure 12: Portions of network status logs on a 12-mile route.

Duration in 3G: In the unestablished call case, we run an 8s-interval UDP data flow for various durations from 1 hour to 10 hours. Figure 11 plots the average duration in 3G with downlink and uplink data flows. Specifically, we measure the interval from the time when the call ends to the instant when the UDP session stops. It is not surprising to see that the phone gets stuck in 3G as long as the data flow is alive (10 hours are observed and there is no sign of limit). Interestingly, we find that the duration in 3G for the downlink case is usually shorter than the uplink case, e.g., we observed 2 hours in 3G networks during a 3-hour data session test. This is because not all UDP downlink packets can be sent successfully under packet loss. The longer the session lasts, the more likely a packet interval goes beyond the transition threshold (15/20 seconds). We also test with TCP flows; Both uplink and downlink flows perform similarly to the UDP uplink case (never expires), because TCP retransmits packets upon losses.

Mobility: We also examine whether a 4G user may go back to LTE networks under mobility where handoffs are triggered. We repeat the above experiment when driving on a 12-mile local route. It takes about 35∼45 minutes. Note that the call ends before driving, i.e., the 4G user gets stuck in 3G before driving starts. In the meantime, we bring another 4G phone without any data session. It is used to collect network status events such as network type and RSSI. Figure 12 plots a portion of the network status logs at two phones over this route via OP-I. The results for OP-II are similar. It is easy to see that, the 4G phone freely switches among 2G, 3G and 4G networks in the absence of data; however, the 4G phone with data only switches among 2G and 3G networks. It never goes back to 4G LTE networks, even though the 4G LTE network signals are stronger than 2G/3G in certain areas, e.g., [11.5, 12].

3G/4G Speed Gap: To quantify the performance impact of being stuck in 3G networks, we measure the speed of 4G LTE and 3G HSPA networks at different hours of a day. We use the SpeedTest tool [6]. Figure 13 plots the average uplink and downlink speed of 3G/4G networks at different times. 4G outperforms 3G in most cases, especially at midnight (with lighter traffic) and for the uplink. For example, 4G users experience 83.4% uplink improvement and 53.8% downlink gain at 11PM. Over all test hours, the average improvement is 70.4% and 31.9% for uplink and downlink, respectively. However, the downlink gap between 3G and 4G shrinks at

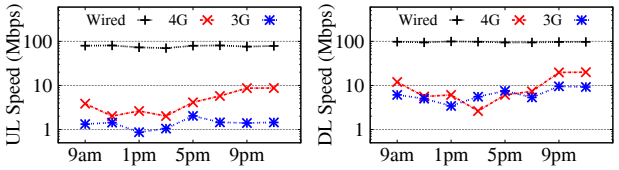


Figure 13: 3G/4G speed at different hours of a day via OP-I (Left: uplink; Right: downlink).

Application	Type	TCP/UDP	Behavior
Webkit	Bursty	TCP	Respond slowly, seldom abort
Gmail	Bursty	TCP	Respond slowly, occasionally abort
Facebook	Interaction	TCP	Respond slowly, seldom abort
AndFtp	Transferring	TCP	Transmit slowly or abort later
Skype	Interaction	UDP	Suspend, abort later
Youtube	Streaming	TCP	Suspend (abort if call unestablished)
PPStream	Streaming	UDP	Suspend (abort if call unestablished)
Pandora	Streaming	TCP	Suspend

Table 3: Application behavior when voice call arrives

other times. To our surprise, 4G performs worse than 3G at 3PM (1.8 Mbps vs. 4 Mbps). This shows that, the deployed LTE network might not achieve what it claims at all times. We also note that, the 3G/4G speed varies with locations (depending on the radio link quality and the traffic load). In general, throughput drop occurs when 4G users get stuck in 2G/3G networks.

6. DATA APPLICATION ABORT

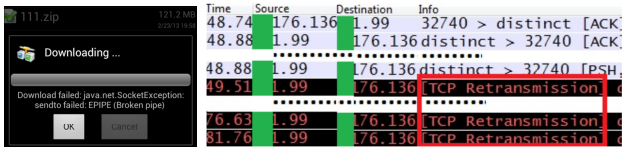
We find that voice calls might even result in data application aborts in operational LTE networks. In this section, we first show real application behaviors under various call operations, such as dialing and answering a call. We then diagnose the root cause for application aborts.

6.1 Popular Applications

We test eight popular mobile applications while receiving a CS voice call. These applications include web browsing (via WebKit), FTP downloading, Gmail, Facebook, Skype voice calls, Youtube, PPStream (P2P video streaming), and Pandora (music playback over radio broadcast). We observe that these applications might behave abnormally when the voice conversation ends on all the phone models for both carriers. Upon voice call completion, five applications except Youtube, PPStream and Pandora might abort, and Pandora suspends for tens of seconds until the playback status turns from "stopped" to "playing." YouTube and PPStream might abort in other calling scenarios.

FTP downloading: In our experiment, a mobile client downloads a 121 MB file from our FTP server. Figures 14(a) and 14(b) show the error dialog at the mobile client and the TCP trace captured by Wireshark at our FTP server. When the voice call ends around the 47th second, the mobile client experiences a socket exception error, i.e., `sendto failed`, in Android OS. File downloading stops afterwards. On the FTP server side, it first attempts to retransmit packets (10 retries over 33 seconds are observed) to the client and finally tears down the TCP connection, since no response is heard from the client. Note that mobile phone aborts earlier at 48s before the server tears down the TCP connection at 82s.

Skype voice calls: When the CS call ends, Skype is designed to resume its original call session; However, it may still fail similarly to FTP, and the root cause will be discussed in Section 6.3. The difference is, during the CS voice call, Skype holds its ongoing call but FTP download keeps going. Skype is not allowed to simultane-



(a) Error at the client (b) Wireshark traces at the server

Figure 14: An example of FTP application abort.

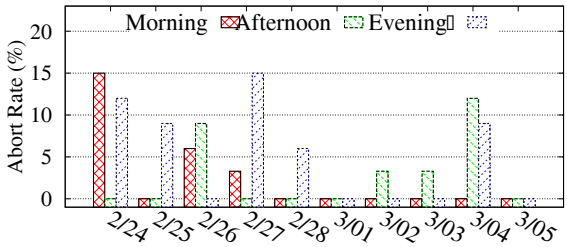


Figure 15: 10-day FTP downloading abort ratio (OP-I).

ously work with CS voice calls due to the conflicting usage of the speaker. So are YouTube, PPStream and Pandora.

Web, Facebook and Gmail: Similar to FTP, they are unable to fetch the attempted web content when they abort upon call completion. It is clearly observed when they fetch big-sized data, e.g., a high-resolution image or an email attachment. For short data sessions (e.g., fetching a html page), the abort takes places but with negligible effect.

Youtube and PPStream: Upon call arrival, both video streaming operations pause. They never automatically resume when the call ends (i.e., a manual replay is required). However, when an incoming call hangs up before reaching the client, Youtube and PPStream might still abort.

Pandora: Similar to Skype call, it is suspended upon the call arrival and resumed automatically after the call ends.Except that the suspension lasts tens of seconds, no other abnormal (i.e., being aborted) behaviors are observed.

Table 3 summarizes these application behaviors, including application aborts and "slow response" due to throughput slump described in Section 4. In fact, application aborts depend on how these applications handle the failure of data sessions, which is triggered by an completed call, in their own manners. The first five applications abort because they do nothing once the original data sessions terminate, whereas Pandora automatically starts a new session once the old one fails. Youtube and PPStream do not take action since they already stop their data sessions once a voice call starts. Note that applications do not abort every time a call completes. We next study when and how often these applications abort.

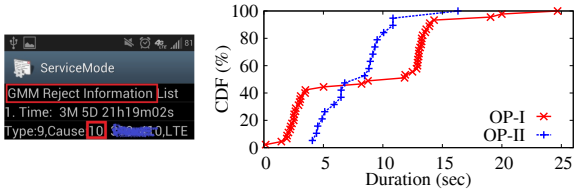
6.2 How Often Application Aborts

We conjecture that these application aborts are caused by voice calls. In our test scenario, a 4G user dials out and hangs up the outgoing call later. In the meantime, we run FTP downloading. The results for other applications are similar. For OP-I, we use Samsung Galaxy S3 and LG Optimus G at two locations (home and campus), during the morning (9am-12pm), the afternoon (1-5pm), and the evening (7-10pm), for 10 days, from February 24 to March 4, 2013. For OP-II, we test Samsung Galaxy S4 and iPhone 5 from June 17–21, 2013. Each test has at least 15 runs. We observe similar abort ratios, independent of the phone model.

Figure 15 plots the 10-day abort percentage for OP-I. The applications do abort but only with certain probability. It confirms that

Seconds	OP	EVENT	TYPE	CID	RSSI	IP
52.84	OP-I	CALL	HANG UP			10.xx.xx.51
53.41	OP-I	NET	UMTS	5****075	-67	10.xx.xx.51
54.30	OP-I	NET	UMTS	5****075	-67	10.xx.xx.51
55.26	Unknown	NET	Unknown	n/a	-113	n/a
56.28	Unknown	NET	Unknown	n/a	-113	n/a
...	...	...	...	...	...	...
69.26	OP-I	NET	LTE	1*****223	-70	10.yy.yy.11

Table 4: Logs of network status at the mobile phone.



(a) Implicit detach (b) CDF of Reattach time

Figure 16: Cause of being kicked out and reattach time.

operational LTE networks are still largely successful. In two worst-case settings (a morning slot and an evening slot), about 15% of tests fail. However, over the 10-day period, the average failure percentages for the morning, afternoon and evening are merely 2.4%, 2.7%, and 5.1%. For OP-II, the average abort ratio is 5.7% in the 5-day test.

6.3 Root Cause: Being Detached

We now explore the root cause for application aborts. We find that it is because the mobile phone is detached from the cellular network when performing an 3G→4G handoff to return to the LTE network after a voice call. Table 4 logs the network status at the phone when an application aborts using OP-I. The call conversation ends at 52.84s. In 2.42 seconds, the user is kicked out of the carrier network (indicated by "unknown"). It also loses its original IP address. This disconnection lasts for about 14 seconds before the 4G phone reconnects to the network. However, upon reconnection, the phone is assigned a brand new IP address. Consequently, the original data session fails and those applications not supporting automatic application-level recovery finally abort.

Another plausible root cause is that power consumption at the UE exceeds the permissible budget when initiating and maintaining radio access bearers for simultaneous voice call and data service. If it were correct, we would observe it when users concurrently access voice and data services, and application abort rate may also depend on phone models. However, our experiments show that this never occurs. Application abort only happens after the call ends, i.e., the UE is being switched back to 4G. The application abort rate observed on different phones is also similar.

To find out why 4G users are kicked outside the cellular network, we enable the service mode of Samsung S3 where low-level cellular network traces can be observed. Figure 16(a) displays the screen snapshot when an application aborts. It states that, the GMM (GPRS Mobility Management) operation [1] is rejected due to an error (cause ID: 10). In this case, an inter-system handoff (from 3G to 4G) request is rejected because it has been implicitly detached [1]. The "Implicit Detached" indicates that the phone is detached by the network without any notification. It typically occurs when the network fails to communicate with the UE. Once this error occurs, the UE has to perform re-attach procedure to associate with carrier networks again [1]. A new IP address will be assigned for OP-I, whereas the same IP address is used for OP-II. However, the NAT (Network Address Translation) mapping for the UE is no longer available. The UE is thus unable to receive packets using the same data session no matter whether the IP address changes.

No packet delivery is allowed until this procedure completes. Figure 16(b) plots the CDF of reattach time. It shows that, 90% of re-attaches would finish within 15 seconds for OP-I, and 95% of re-attaches is shorter than 11s for OP-II.

We are not sure why the network detaches the user when a CSFB call ends, due to lack of status information inside the network. It might be caused by the failure of inter-system (3G to 4G) handoff, due to insufficient resources (resource occupied by CSFB) or unsynchronized user information between 3G and 4G [1].

7. REVERSE IMPACT: MISSED CALLS

We next show how the PS data service may adversely affect CS voice calls. We find that 4G LTE users may miss their voice calls while starting PS data access. When the caller makes a call but the callee starts PS data access almost simultaneously, the caller hears success signals (e.g., alerting tone) so that he/she believes that the call has been made but is not answered. In the meantime, the callee receives no incoming-call request (e.g., no ringing or vibrating). Everything else operates normally, but the callee is unaware of missing a call.

We test it with two experiments. First, we make a call while the callee starts to turn on its PS data network (i.e., network attach). In all test runs (> 20), all calls have been missed. Second, we make a call when the data network is either off or already on, i.e., the callee does not turn on his/her PS data network while caller makes call; In this case, all calls have succeeded. The same results are observed on all phones for both carriers. We note that, in case of missing calls, the caller may have an option to leave a message if his/her voice-mailbox feature is enabled. The voice-mailbox feature is free in the US, so the adverse effect of missing calls can be greatly relieved. However, not all operators support free voice-mailbox features, so missed calls may incur inconvenience.

Root cause: We analyze the *NetworkStatus* trace logged on the callee's phone. With an incoming call request, the callee is implicitly detached by the network (same as Section 6). During the period (before network re-attach completes), the mobile loses connectivity with carrier networks. We next seek to understand why the caller hears an alerting tone, thus misinterpreting that the call has been established.

We examine the incoming CSFB call flow of Figure 3. In Step 2 (Paging), the MSC pages the UE through MME. Following the *mobile terminated call procedure* [1] in the CSFB standard, the UE will respond with `Service Request` [1] to the MSC, then the MSC sends an indication (i.e., an alerting tone) to the caller. In fact, it happens before the handoff to 3G networks occurs; The caller is acknowledged no matter whether the callee is kicked out of carrier networks or fails to handover to 2G/3G networks. This results in *asynchronous* call status at the caller and the callee. This scenario differs from the call establishment process in 3G networks, where the caller hears the alerting tone only after the callee is found by the network via paging. We believe that it is a fundamental issue in cellular networks, rather than a operator-specific implementation choice (despite observed in both operators). PS data and CS voice are performed independently on their data plane, but share common network states on the control plane. Imprudent control in one domain may impose unexpected impact on the other domain.

8. SOLUTION

We now describe our solution to mitigating the negative impact incurred by CSFB voice calls in 4G LTE networks. We use a combination of techniques to address all four issues: performance degradation of TCP-based data sessions in the presence of CSFB calls, unexpected application abort, lost 4G connectivity, and missed calls during PS service.

Mitigating TCP performance degradation We first note that the TCP issue is due to inter-system handoffs; it cannot be fixed from the CSFB protocol itself since the handoffs are a fundamental feature of CSFB. We follow the popular middlebox-based approach [30] in our solution. Our scheme differs from the related work [12,22,24] in that we focus on voice-triggered handoffs rather than mobility-induced handoffs. We split the TCP connection into two sessions: one between the middlebox and the application server, and the other between the middlebox and the UE. The UE detects handoffs induced by CSFB, and sends suspension request to the middlebox. Upon receiving a suspension request, the middlebox freezes its retransmission timer and caches data packets from the application server for about 15 seconds. The parameter is chosen because 90% of data suspension time is less than 15 seconds (Figure 16(b)) in both handoff and application-abort cases. Once receiving a resumption request from the UE after the handoff completes, the middlebox immediately retransmits its cached packets to the UE. Note that the timeout value in CSFB is decided by the UE, rather than the operator, thus different from the mobility-induced handoff case. Our prototype implementation on Android phones and the middlebox proxy shows that, our solution is 2-50% faster in packet reception recovery than the standard TCP, and the average improvement is 18%. The main merit of our solution is that it can be readily integrated with existing carrier middleboxes, but it also incurs more complexity and handles only the TCP flows.

Handling lost 4G connectivity In our solution, we let the carrier initiate the inter-system handoff (i.e., 3G $\rightarrow$ 4G) to switch the user back to 4G when both conditions are met: (1) no ongoing voice call exists; (2)the duration the user stays in the 3G network is longer than certain threshold (e.g., 60 seconds). Our scheme not only addresses the issue of lost 4G connectivity, but also avoids unnecessary handoffs. The downside is that, the carrier has to maintain a timer for each 4G user to record how long (s)he stays in 3G. In contrast, another possible solution is that the operator immediately switches the user back to the 4G LTE network once the call completes. However, it may lead to more CSFB-induced handoffs since the caller may redial for incomplete call conversation (e.g., the Operator-I's scenario). Moreover, it may lead to a potential security loophole that incurs significant inter-system handoffs on the 4G user via repetitively dialing and hanging calls up from the malicious user.

Handling application abort This issue can be addressed by either an in-network approach (e.g., following Operator-II's IP assignment policy that assigns the same IP address to the UE and keeps the NAT mappings after the UE reattaches to carrier networks), or an out-of-network approach (e.g., via the middlebox).

Our solution is still based on the middlebox. We borrow ideas from Cisco AnyConnect [4], which offers a mobile VPN scheme that allows for the UE to reconnect its VPN server via different IP addresses to proceed the session established earlier. We do not require secure connections that encrypt each transmitted packet, but enable the mobile device to connect to the proxy server with a different IP address. The UE is thus able to resume data sessions that are established earlier between the middlebox and the application server. Our scheme thus saves computing power and energy consumption at the mobile device. The downside is that the middlebox may pose as the bottleneck for the transmission rate to the UE.

Handling missed call Our solution slightly modifies the current CSFB specification. Upon receiving `Service Request`

(introduced in Section 7), the MSC does not send an indication of user alert to the caller. This notification is deferred until `Call Setup Response` arrives at the MSC. Consequently, Alice will not hear the alert tone before Bob successfully hands over to the 3G network and enters `Call_Received` state. Our tests show that, in the current practice of 4G LTE CSFB, Alice hears the alert tone about one second before Bob's phone rings. Therefore, our solution may increase only one second based on our estimate to establish the voice call when the caller hears the alert tone. We thus address the issue with little cost (about 1-second extra waiting time). The downside is that our proposal requires modifications on the CSFB specification.

9. DISCUSSION

We now discuss two solution alternatives to CSFB to support voice calls over 4G LTE.

CSFB: An interim solution to VoLTE or not? Some people may argue that, CSFB is only an interim voice solution, and all its problems will disappear once the ultimate solution of VoLTE is deployed. However, the global deployment of VoLTE is not foreseen in the near future. VoLTE is only supported by three small operators so far, MetroPCS (US), SK Telecom and LG Uplus (both in South Korea) [5]. Most operators, including four of the top-five global operators, China Mobile, Vodafone, Bharti Airtel, and Telefonica, make plans to deploy or have deployed CSFB as their voice solution in 4G LTE networks [2, 3, 7, 8].

In fact, VoLTE has several drawbacks that may impede its large-scale deployment. They include high technology complexity, challenges to ensure guaranteed service for voice, and more energy consumption of mobile clients. VoLTE is purely PS based, and offers voice service via its IMS (IP Multimedia Subsystem) [1]. The first version of IMS was released in 2002. During the past decade, operators have little incentive to deploy IMS, because of its high cost in deployment and maintenance, as well as its operation complexity [16]. Moreover, it poses challenges to deliver guaranteed service for voice calls on top of the IP-based best-effort service. Fundamentally, it is the old, challenging Internet QoS problem, and cellular networks need extra mechanisms to ensure carrier-grade quality when migrating CS voice calls into the PS domain. Last, the client may consume more power when using VoLTE. We run two experiments to estimate the power consumption via VoLTE. The tests do not intend to be very accurate but show us some hints on the energy aspect of VoLTE. One experiment is to make a CS voice call only, and the other is to deliver a 12.2 Kbps data flow that emulates the traffic of a voice call in VoLTE [18]. We measure the power consumption of the client device while its LCD screen is off. The CS call takes 0.96 W while the PS data service takes 1.42 W. Another measurement conducted in [17] shows that the energy consumption of a VoLTE call can be twice as much as that of a 2G call. Therefore, VoLTE may pose potential energy issues for the client device. Moreover, to retain the capability of being paged for an incoming call, the mobile device may have to power on its data network interface all the time but not on demand.

SVLTE: Alternative to CSFB? SVLTE (Simultaneous Voice and LTE) offers an alternative solution to voice call in 4G LTE. It allows for a phone to simultaneously use both networks, thus avoiding the issues raised by CSFB. However, this technique is only applied to CDMA 1xRTT [29]; it is not compatible with LTE plus 3G UMTS systems that are widely deployed. More importantly, it requires *two radios*, one for PS and one for CS. Therefore, power consumption may become a concern. We measure the power consumption of a SVLTE-capable phone (Samsung Stratosphere). Upon answering a voice call with/without background data service (50 Kbps), the phone consumes power at about 1.0 W and 2.39 W, respectively. When both radios are on to support both CS and PS, the phone consumes extra 0.97 W than VoLTE. This issue has been reported by [15] and our measurement results are similar.

10. RELATED WORK

Cellular networks have been an active research area in recent years. However, the interplay of voice calls and data services in 4G LTE networks has remained largely unaddressed in the research community.

How to better support voice calls in LTE networks has appeared in the literature [9, 14, 20, 21, 23]. All such studies seek to improve the voice service quality, but do not study the potential mutual impact of voice and data, which is the focus of this work. Various performance aspects of 2G/3G/4G networks have been studied, and new solutions have been proposed, e.g., [11, 19, 22, 24–26, 28]. These studies mainly focus on wireless data, whereas we examine interactions between voice and data.

Our proposed solution fix also bears similarity to existing designs. Our goal is to address the four identified new issues. We thus borrow several ideas freely from the literature and do not claim novelty. The general middlebox-based solution is a popular industry practice [30]. How to improve TCP under mobility-triggered handoffs has been well documented in early papers, e.g., [22, 24], though our handoff events are induced by CSFB voice calls.

11. CONCLUSION

With the success of the Internet, the 4G LTE technology has also adopted packet-switched (PS) delivery while abandoning the circuit-switched (CS) model. The PS service facilitates mobile data but causes problems for voice calls, which required guaranteed (carrier-grade) service. Given the undisputed importance of both data and voice, LTE carriers have used CSFB as a popular interim solution to CS voice. In this work, we use experiments to study how the CSFB-enabled voice interacts with the PS-based data in operational LTE networks. To our surprise, voice and data indeed interfere with each other. Voice may cause data to reduce throughput, abort applications, and lose 4G connectivity. Data may also cause the voice service to miss incoming calls.

We believe that the identified issues lie in both the design of the CSFB technology and its engineering implementation. The key features, including the finite state machine, the inter-dependency between data and signaling, and the third-party triggered handoff, are all fundamental to CSFB. Therefore, the problems stem from the design of CSFB. They include (1) deadlocked state transitions in the finite state machine of RRC, (2) unexpected coupling between signaling and data; and (3) arbitrary triggering of handoffs by a third party without security protection. The fundamental problem is that, users demand both data and voice services on the LTE phone device. Voice and data thus interact with each other since both use the shared radio. While the support for PS-based data tends to be simple, the signaling and control operations to enable CS-based voice are complex. Throughout the life cycle of a voice call, any procedure may go wrong. Consequently, any failure or exception in the process may affect both voice and data through the phone. The goal of this work is to better understand such systems by exposing some of such cases and devising solution fix at the early stage of global LTE deployment.

Our somewhat biased results should not be interpreted as the common failures of operational LTE networks with CSFB. It remains largely successful in practice. Some may argue that the

problem will be short lived, since CSFB is only an interim solution. Therefore, all identified problems would disappear once the ultimate VoLTE solution is deployed. However, current practice shows that VoLTE does not look promising within the 3–5 year horizon. Most operators, including four of the top-five worldwide operators have committed to CSFB. Moreover, another alternative SVLTE also has its own drawbacks. Given the growing interest on CSFB for LTE, we seek to locate the problem and find its solution.

12. ACKNOWLEDGMENTS

We greatly appreciate the insightful and constructive comments from our shepherd, Dr. Li Erran Li, and the anonymous reviewers. We also thank Zhe Wen and Xingyu Ma for their contributions at the early stage of this work and all participants in the experiments.

13. REFERENCES

[1] 3GPP Specification: TS23.060, TS23.401, TS23.228, TS23.272, TS24.008, TS36.331, TS36.304. http://www.3gpp.org.

[2] Bharti Airtel to offer voice services to LTE customers via GSM #MWC13. http://www.nokiasiemensnetworks.com.

[3] China Mobile selects Nokia Siemens Networks for large, multi-city, TD-LTE deployment. http://www.nokiasiemensnetworks.com.

[4] Cicso AnyConnect Secure Mobility Client. http://www.cisco.com.

[5] MetroPCS, SK Telecom, LG Uplus Claim VoLTE First. http://www.telecomengine.com.

[6] Speedtest.net - Ookla. http://www.SpeedTest.net.

[7] Telefonica Germany achieves to transfer a phone call from LTE to UMTS without interruptions. http://blogthinkbig.com.

[8] Vodafone Supports CSFB. http://www.webwire.com/ViewPressRel.asp?aId=170837.

[9] Voice over LTE. http://www.gsma.com/technicalprojects/volte.

[10] iphone 5 review, 2013. http://www.anandtech.com/show/6330/the-iphone-5-review/18.

[11] P. K. Athivarapu, R. Bhagwan, S. Guha, V. Navda, and et.al. Radiojockey: mining program execution to optimize cellular radio usage. In *ACM MobiCom*, Aug. 2012.

[12] H. Balakrishnan, V. N. Padmanabhan, S. Seshan, and R. H. Katz. A Comparison of Mechanisms for Improving TCP Performance over Wireless Link. In *ACM SIGCOMM*, 1996.

[13] CNET. Competitive wireless carriers take on at&t and verizon, 2012. http://news.cnet.com/8301-1035_3-57505803-94/competitive-wireless-carriers-take-on-at-t-and-verizon.

[14] A. Dhananjay, A. Sharma, M. Paik, J. Chen, and et.al. Hermes: Data Transmission over Unknown Voice Channels. In *MobiCom*, 2010.

[15] B. Ekelund. LTE, Telephony and Battery Life, 2012. `http://www.stericsson.com/technologies/VoLTE_power_consumption_white_paper_R6.pdf`.

[16] Ericsson. Current status of IMS deployment strategies & future developments, 2012. http://www.itu.int/.

[17] K. Fitchard. Voice calls over 4G LTE networks are battery killers, 2012. http://gigaom.com/2012/11/28/volte-calls-consumer-twice-the-power-of-2g-voice-calls/.

[18] GSMA. IR.92: IMS Profile for Voice and SMS, Mar. 2012.

[19] J. Huang, F. Qian, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck. A Close Examination of Performance and Power Characteristics of 4G LTE Networks. In *ACM MobiSys*, June 2012.

[20] Y. Jouihri and Z. Guennoun. Best selection for operators starting lte deployment towards voice services. In *IEEE ICMCS*, May 2012.

[21] M. Keeley. Deployment Challenges Await In VoLTE QoS User Equipment, 2012. http://mobiledevdesign.com/tutorials/deployment-challenges-await-in-volte-qos-user-equipment-1210/.

[22] X. Liu, M. Seshadri, A. Sridharan, H. Zang, and S. Machiraju. Experiences in a 3G Network: Interplay between the Wireless Channel and Applications. In *ACM MobiCom*, Sep. 2008.

[23] J. Namakoye and R. Van Olst. Performance evaluation of a voice call handover scheme between LTE and UMTS. In *AFRICON*, 2011.

[24] C. Paasch, G. Detal, F. Duchene, C. Raiciu, and et.al. Exploring Mobile/WiFi Handover with Multipath TCP. In *CellNet*, Aug. 2012.

[25] C. Peng, G. hua Tu, C. yu Li, and S. Lu. Can We Pay for What We Get in 3G Data Access? In *ACM MobiCom*, Aug. 2012.

[26] C. Peng, C.-y. Li, G.-H. Tu, S. Lu, and L. Zhang. Mobile Data Charging: New Attacks and Countermeasures. In *CCS*, Oct. 2012.

[27] F. Qian, Z. Wang, A. Gerber, Z. M. Mao, and et.al. Characterizing Radio Resource Allocation for 3G Networks. In *IMC*, Nov. 2010.

[28] G.-H. Tu, C. Peng, C.-Y. Li, X. Ma, H. Wang, T. Wang, and S. Lu. Accounting for roaming users on mobile data access: Issues and root causes. In *ACM MobiSys*, June 2013.

[29] A. D. V. Vanghi and B. Vojcic. *The cdma2000 System for Mobile Communications: 3G Wireless Evolution*. Pearson Education, 2004.

[30] Z. Wang, Z. Qian, Q. Xu, Z. Mao, and M. Zhang. An Untold Story of Middleboxes in Cellular Networks. In *ACM SigComm*, Aug. 2011.

APPENDIX

A. CASE ANALYSIS OF 3G DURATION

We use case study to analyze 3G durations and validate how each handoff rule take effects in each case of Figure 10(b). In particular, we will see how two timers and intra-handoff events affect the time back to 4G.

Case (I): 1B/10s The first packet packet delivery triggers an intra-3G handoff (Rule 4) and then it enters into `DCH/FACH` (see ① in the plot). Since the first intra-3G handoff usually takes a longer time (about 8-10 seconds), the second packet (at 10th second) is delivered via 3G since it is still in `DCH/FACH`. In this case, the packet interval is very close to T_{idle} (also 10s); Thus, the 3G duration is sensitive to whether the packet arrives before T_{idle} times out. In this example, the third packet arrives slightly earlier before RRC becomes `IDLE`, whereas the fourth packet (at the 30th second) arrives after that (see ②). Therefore, at 30th second, the timer $T_{3G\to 4G}$ is set and the handoff to 4G is triggered 5 seconds later. Finally, it returns to 4G at 37th second (i.e., the handoff takes about 2 seconds).

In the best case, the handoff timer is triggered just before the 3rd packet, and the duration is about $(20 + 7) = 27$ seconds. We observe that the duration varies dramatically due to its time sensitivity and goes up to 120 seconds. The detailed trace analysis shows the fourth packet does not trigger an intra-3G handoff. We find that it is because starting the timer $T_{3G\to 4G}$ happens almost at the same time as the request to trigger an intra-3G handoff. We gauge that, the handoff to 4G has higher priority than an intra-3G handoff and thus the intra-3G handoff is dismissed. We attribute this rule to the operator-specific implementation. It has been validated while we slightly increase the interval. The intra-3G handoff is triggered if the interval is larger than 10s, or in [11s, 14s]. Hence $T_{3G\to 4G}$ is reset and the phone never returns to 4G.

Case (II) 1KB/10s It never returns to 4G since the packet interval is smaller than 15 ($T_{idle} + T_{3G-HO} \approx 10 + 5$) seconds. It stays in `FACH/DCH` state and each packet transmission triggers an intra-3G handoff (see ③ in the plot). It also explains why the phone never returns to 4G when the 1KB-packet interval is smaller than 15 seconds.

Case (III) 1B/15s The difference from Case (I) is, before the third packet arrives, RRC turns into `IDLE` at about 25th second (10 seconds after the 2rd packet delivery) and thus the timer $T_{3G\to 4G}$ is set (Rule 1). It is also sensitive whether $T_{3G\to 4G}$ times out before the next packet arrival. If so, it triggers an handoff back to 4G, as shown in Figure 10(b). Otherwise, $T_{3G\to 4G}$ is reset and the duration in 3G is prolonged. This is why the duration also fluctuates around 15th second. When we slightly increase the interval (e.g., 15.5 seconds), the timer $T_{3G\to 4G}$ times out always before the next packet arrival. It returns to 4G similar to all the $> 15s$ case.

Case (IV) 1KB/15s Here, it is sensitive to how long the intra-3G handoff lasts. If the intra-3G handoff finishes within 5 seconds (e.g., 4.5 seconds), RRC turns `IDLE` and $T_{3G\to 4G}$ is set before next packet arrival. Otherwise, the triggered intra-3G handoff is to reset the $3G \to 4G$ handoff timer. Similarly, we induce that 20s is another transition interval for 1KB packets. For those intervals in [16s, 19s], the packet interval is larger than 15 seconds ($T_{3G-HO} + T_{idle}$), but is not large enough to wait for $T_{3G\to 4G}$ timeout. The subsequent intra-3G handoff resets this timer, so it will never return to 4G. It is also similar to the 1B-packet case with the interval in [11s, 14s].

In addition, it also explains the duration remaining in 3G without ongoing data. The timer 3G→4G handoff is set at the start, and the handoff is usually triggered 5 seconds later when it falls back to 3G. The handoff takes another 2–3 seconds to finish. This is why it is smaller than the one with data (at least first two packets are sent in 3G).