Performance Characterization and Call Reliability Diagnosis Support for Voice over LTE

Yunhan Jack Jia, Qi Alfred Chen, and Z. Morley Mao
University of Michigan
{jackjia, alfchen, zmao}@umich.edu

Jie Hui, Kranthi Sontineni, Alex Yoon, Samson Kwong, Kevin Lau
T-Mobile USA Inc.[1]
{jie.hui, kranthi.sontineni1, alex.yoon4, samson.kwong, kevin.lau}@t-mobile.com

ABSTRACT

To understand VoLTE performance in a commercial deployment, in this paper we conduct the first comprehensive performance characterization of commercially deployed VoLTE, and compare with legacy call and over-the-top (OTT) VoIP call. We confirm that VoLTE excels in most metrics such as audio quality, but its call reliability still lags behind legacy call for all the three major U.S. operators.

We propose an on-device VoLTE problem detection tool, which can capture new types of problems concerning audio quality with high accuracy and minimum overhead, and perform stress testing on VoLTE call's reliability. We discover 3 instances of problems in the early deployment of VoLTE lying in the protocol design and implementation. Although the identified problems are all concerned with the immature LTE coverage in the current deployment, we find that they can cause serious impairment on user experience and are urgent to be solved in the developing stage. For example, one such instance can lead to up to 50-second-long muting problem during a VoLTE call! We perform in-depth cross-layer analysis and find that the causes are rooted in the lack of coordination among protocols designed for different purposes, and invalid assumptions made by protocols used in existing infrastructure when integrated with VoLTE. We summarize learnt lessons and suggest solutions.

Categories and Subject Descriptors

C.2.1 [**Network Architecture and Design**]: Wireless communication; C.4 [**Performance of Systems**]: Measurement techniques, performance attributes

Keywords

VoLTE; VoIP; QoE; Trouble Shooting Cellular Network

[1]The views presented in this paper are as individuals and do not necessarily reflect any position of T-Mobile.

MobiCom'15, September 7–11, 2015, Paris, France.

DOI: http://dx.doi.org/10.1145/2789168.2790095.

1. INTRODUCTION

For reasons including improved cost effectiveness, spectrum use, and service quality, wireless operators are actively deploying VoLTE (voice over LTE), to support voice as another application on the data network, similar to over-the-top (OTT) VoIP applications. To ensure high service quality to end users, VoLTE relies on specialized support in both end devices and network architecture [38, 44], as well as dedicated radio resources during the voice call, to achieve an ambitious goal of completely replacing the legacy call in the long term.

VoLTE deployment by U.S. operators is still at an early stage, but it is important to empirically understand its performance from the perspective of mobile end users, especially compared to legacy call and OTT VoIP call. In this paper, we present the first systematic study using a variety of important call performance indicators for the VoLTE service provided by three major U.S. carriers, and also compare VoLTE with legacy 3G call, Skype, and Google Hangouts Voice call using controlled experiment. Our results show that VoLTE is superior in a majority of the metrics: it delivers the best audio quality, much better than legacy 3G call; compared to OTT VoIP call, it consumes 83% less data, 75% less energy, and also has around 40% shorter call setup time.

The major challenge of VoLTE in the early deployment is the inadequate LTE coverage compared with 2G/3G, which affects VoLTE call's reliability under undesired network condition and in the mobility case. To address the challenge in the transition phase, various supporting techniques are adopted by VoLTE service providers to reinforce the reliability (detailed in §2). However, Our study reveals that in the current VoLTE deployment, call reliability is still not competitive with legacy calls. Call failure ratio, including setup failure and unintended call drop, is almost 5× higher (§4.1) than the legacy call, which is a known problem to severely impair user experience [16] and is a common problem seen in all the three major US carriers' network. It is clearly imperative for operators to address the VoLTE call reliability challenges as soon as possible.

Unfortunately, systematically capturing and diagnosing VoLTE problems remains rather challenging, especially for new problems such as audio muting [46]. Current mechanism adopted by carriers to capture them largely depends on user feedback [47], which is not only unreliable, but also insufficient.

In view of these challenges, in this paper, we pursue two directions to identify the root causes of these prevalent VoLTE problems. First, we design and implement a VoLTE problem detection tool (§4.2) to capture audio quality issues of VoLTE with low false positive rate (0.65%) and false negative rate (5%). Second, we use stress testing (§5) to uncover VoLTE problems in controlled settings and diagnose the root causes. Three types of problems of the

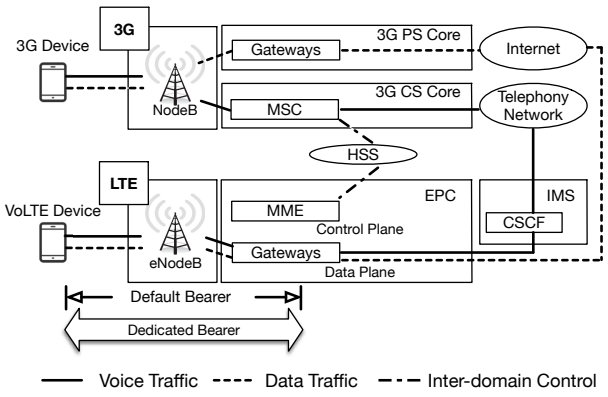

Figure 1: 3G and LTE architecture with VoLTE enabled.

deployed VoLTE service are identified: (1) a lack of coordination between device-originated and network-originated events (§6.1), which leads to consistent VoLTE call setup failure in some cases; (2) incorrectly ordered inter-dependent actions (§6.2), which lead to a high VoLTE call drop rate; (3) a lack of coordination in cross-layer interactions (§6.3), which results in extremely long muting during VoLTE calls. Although we identify that these problems are all temporary in the early deployment, we remind the commity that such problems may always exist as long as we are in a heterogeneous network environment. We summarize learnt lessons in §6 and provide the discussion in §7.

The contributions of this paper can be summarized as:

- We perform the first in-depth systematic study of VoLTE performance in three major U.S. carriers' network, and a comprehensive comparison among VoLTE and related telephony technologies.
- We conduct an extended study of call reliability of VoLTE, together with OTT VoIP and legacy call, and devise a novel online audio quality monitor to discover VoLTE call reliability problems in the production network, which overcomes the challenge of capturing the QoE problems of telephony services in real time.
- We perform stress testing to help identify the root causes for three types of common VoLTE problems, and uncover three instances of lacking coordination within VoLTE related protocols' design and implementation. We perform in-depth cross-layer analysis to deduce the root causes, and suggest potential solutions.

2. BACKGROUND

LTE (Long Term Evolution) is the next generation mobile technology that aims at supporting all services including voice, using a common IP infrastructure without separated voice channels. The architecture of LTE together with the co-existed 3G is shown in Fig. 1. The 3G voice domain, circuit switched (CS) core, transports legacy 3G calls in the same way as a traditional fixed-line telecommunication system. The 4G LTE domain, the packet switched (PS) core, transports data streams between the user and external packet data network such as the Internet via gateways.

2.1 Voice over LTE

VoLTE is technology that allows carriers to transmit voice calls over LTE network and through their IP Multimedia Subsystem (IMS) cores controlled by Call Session Control Functions (CSCF). This means voice calls and data sessions will travel side-by-side over LTE and exchange with Internet through the gateways. During a VoLTE call, VoLTE maintains both control and data sessions. The signaling messages to initiate the call flow over the control plane and use Session Initiation Protocol (SIP) to control the call. Voice packets are encoded with Adaptive Multi-Rate wide band (AMR-WB) codec and are transmitted with RTP protocol over the data plane.

VoLTE is supposed to benefit both carriers and end users. For carriers, VoLTE is expected to be more efficient from both network and spectrum perspective [33], and allows convenient global roaming. For end users, the most noticeable benefit of VoLTE is expected to be the improved voice quality, since it adopts wide band codec, which should make a call clearer than a legacy cell phone call. Furthermore, VoLTE is also expected to have better performance since QoS is guaranteed with dedicated bearer used. As shown in Fig. 1, a bearer is a "pipe line" for transporting data among device, ENodeB, gateways, and other entities. Prior to VoLTE, all the application data is transmitted by a default bearer where QoS is not guaranteed.

Since voice call is the most basic application of a mobile device, users expect VoLTE to provide good audio experience, at least comparable to legacy calls. However, the performance and reliability of VoLTE highly depend on the deployment of the LTE network, which currently is not as ubiquitous as the 2G/3G network. On top of that, to deploy VoLTE, additional complexity added in both the network infrastructure (e.g., IMS) and the process of making a audio call (e.g., CSFB, SRVCC) is very likely to introduce new problems in practice. In view of these challenges, we are motivated to perform the first comprehensive performance characterization of commercially deployed VoLTE, and also diagnose the discovered problems to help carriers improve VoLTE.

2.2 Key Terminology

Legacy call in this paper stands for the 3G/UMTS circuit switch call, which is used before LTE is deployed.

CS fallback (CSFB) is an alternative of VoLTE before the new IMS-based VoLTE architecture is deployed, which redirects a device registered on the LTE network to the 2G/3G network (i.e. fallback) prior to originating or receiving a voice call. After VoLTE is deployed, it is still kept as a standby solution for some carriers: when VoLTE call fails to setup, CSFB will be originated to reestablish the call in CS domain.

Inter Radio Access Technology (RAT) handover stands for a handover between different RAT, which can be either LTE, HSPA+[1], CDMA or other wireless technologies. In this paper, we mainly focus on LTE to HSPA+ handover, which usually happens when leaving LTE coverage, and denote it as *inter-RAT handover* in later sections. By default, an inter-RAT handover will terminate the ongoing VoLTE session if no call continuity technology such as SRVCC (explained below) is supported.

Single Radio Voice Call Continuity (SRVCC) is an advanced LTE functionality that improves inter-RAT handover in terms of VoLTE call continuity. It allows VoLTE call to be seamlessly moved from LTE PS domain to legacy CS voice domain without dropping the call. Differing from CSFB that does handover for call setup process, SRVCC does handover for ongoing VoLTE sessions.

[1]HSPA+ stands for Evolved High-Speed Packet Access. Usually marketed as 4G

3. VOLTE PERFORMANCE CHARACTERIZATION

To validate the performance gain promised by VoLTE in real deployment, in this section, we present our measurement study to characterize VoLTE performance and compare it with its alternative telephony technologies, which are legacy circuit-switched call and OTT VoIP. Note that our focus is on telephony services over cellular network, so other telephony solutions such as Wi-Fi calling [1] are not included.

Table 1 summarizes the results and lists the key performance indicators (KPIs) for a voice call application that are considered in our measurement study, along with their corresponding end-user experience. We cover performance metrics in this section and focus on call reliability metrics in §4.

Experiment setup. We test VoLTE services from three major US carriers, denoted as OP-I, OP-II, and OP-III, which together take 83% of the market share of US. The tesing environment is in good network condition for all the three carriers, which has a LTE Reference Signal Received Power (RSRP) around -95 dBm (5 out of 5 signal strength bars). We also cover the performances in other scenarios, such as undesired network environment and mobility case by simulating network condition and performing drive test.

Based on our measurement results, we find that VoLTE performances of the three carriers are similar in terms of audio quality, network resource requirement, and power efficiency. So in later sections, if not specified, we use OP-I's result to represent VoLTE's general performance and denote as "VoLTE". For comparison, Skype and Hangouts Voice are tested using the same device and data plan as we do for OP-I's VoLTE, and legacy call is also tested in OP-I's network.

Since VoLTE-supported device models vary in different carriers, we use Samsung Galaxy Note 3, LG G3 Vigor, and Samsung Galaxy S5 respectively to test VoLTE of the three carriers. For legacy calls, we use Nexus 5 devices, which doesn't support VoLTE in OP-I's network and will switch back to 3G to complete the legacy CS call. In our experiments the Skype version is 5.0.0 and the Hangouts version is 2.4, which are both the latest versions up to the time we write the paper.

3.1 Audio Quality under Different RSRP

Methodology. To measure the audio quality under different signal strengths, we use a programmable attenuator [27] together with a RF shield box [21] to vary the received RSRP of a mobile device. The RF shield box isolates the devices inside from the over-the-air network signal outside, so that the device can only receive signal from a single antenna connecting the box to the external network. The programmable attenuator is attached to the antenna on the RF box and can attenuate the signal strength of the inside environment within the range of [-95 dBm, -140 dBm].

We use Spirent Nomad HD equipment [45] to calculate the MOS for different applications, which provides us with objective MOS by characterizing the one-way degradation of audio quality. It takes the input audio from microphone on one device, together with the output audio from the earphone jack of another device for comparison, and calculate the MOS based on several metrics including one-way delay and missing frames. The objective MOS given by the equipment is on a 0–5 scale, where 4 and above is regarded as good or excellent quality with imperceptible impairment but below 3 is considered poor quality with annoying impairment [28]. We perform test calls with a 30-minute duration for each application under each RSRP level and the MOS is calculated every 10 seconds. We use uplink MOS to present the result since downlink MOS shows the similar trend according to our experiment.

MOS	OP-II VoLTE	OP-II legacy call
OP-I VoLTE	3.16	3.46
OP-I legacy call	3.28	3.25

Table 2: Median MOS when making calls from device indicated in row to device indicated in column.

Result. Fig. 2 shows the uplink MOS calculated for VoLTE and VoIP applications under different RSRP. Between Skype and VoLTE, Skype has slightly better audio quality over VoLTE in average but larger variance under poor signal strength. We discuss the reason for this in §3.3: Skype tends to use much more bandwidth to ensure a better audio quality. VoLTE and Skype all deliver satisfactory audio quality (mean MOS of 3.8 and above) under reasonably good signal strength levels, while Hangouts Voice lags a little behind. VoLTE demonstrates significant improvement compared with legacy calls, which can only reach 3.3 even in best case.

However, we identify that VoLTE doesn't guarantee superior audio quality all the time considering the lack of cooperation among VoLTE service providers in the current stage of VoLTE deployment. Table 2 shows the mean uplink MOS when making inter-operator calls. Surprisingly, we find that the audio quality of VoLTE call between OP-I and OP-II are even worse than that of the legacy call between them. The cause is that there are two times of transcoding performed by proxies in the network during a VoLTE call between OP-I and OP-II's network [4], which transcode the audio packets to analog and then back to packets, which degrades the audio quality from audio source.

3.2 Audio Quality with Contending Traffic

Methodology. Considering the scenario where applications generate traffic in the background during a voice call session, we are motivated to study the impact of contending traffic on VoLTE and OTT VoIP. We develop a client/server traffic generator with the client running on the testing device, which can control the UDP traffic generating rate of both uplink and downlink. The impact of background traffic on MOS is then measured by tuning the traffic rate during an ongoing call. We only present the uplink result in Fig. 3 since we observe in our experiment that the impact of downlink traffic on downlink MOS is equivalent to the impact of uplink traffic on uplink MOS.

Result. Fig. 3 shows the effect of background upload traffic on audio quality. We confirm that VoLTE won't be affected by background traffic due to its intrinsic dedicated bearer support, which ensures bandwidth isolation between voice and data. The MOS of Skype, on the other hand, degrades quickly to the level of "Annoying impairment" with increasing upload traffic rate, which may be caused by the buffering on the device and in the network.

3.3 Network Resource Consumption

Methodology. We study the network resource consumption of different voice call applications by measuring their throughput when delivering the same speech sample. Considering the bit rate adaptation of the codec, we conduct experiment in good signal strength level and capture the throughput in 3 scenarios that may present in a call session, including silent, speech, and mute, where speech is when the person is talking, silence is when the person is just listening with only background noise, and mute is when the person mutes himself by clicking the mute button during the call. Using Qualcom eXtensible Diagnostic Monitor (QXDM) [39], which can collect radio layer information on the device, we are able to study

§	Figures & Tables	Comparison entities	End user experience	Key performance indicators
3.1	Fig. 2, Table 2	V,S,H,L	Speech quality during a call	Audio quality under various signal strengths
3.2	Fig. 3	V,S,H		Audio quality with contending traffic
3.3	Fig. 4	V,S,H	Data charge [1]	Network resource requirements
3.4	Fig. 5	V,S,H	Smooth audio experience	Jitter, and one-way delay
3.5	Fig. 6	V,S,H,L	Time it takes to start ringing	End-to-end call setup time
3.6	Table 3	V,S,H,L	Battery life	Power efficiency
4.1	Fig. 7	V,S,H,L	Ability to make and maintain phone call	Call reliability

[1] Some carriers charge VoLTE by call durations instead of data usage.

Table 1: Summary of measurement results in §3. **V**, **S**, **H**, and **L** stand for VoLTE, Skype, Hangouts Voice, and Legacy call.

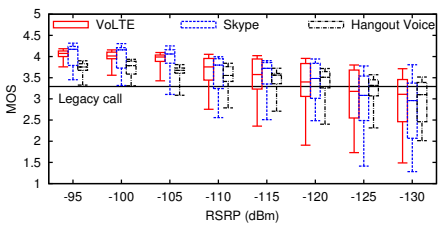

Figure 2: Uplink MOS under variant signal strengths. The dotted line represents the MOS a legacy call achieves in best case.

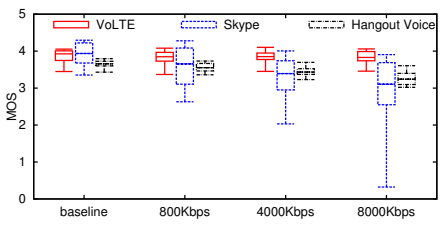

Figure 3: Uplink MOS under different background uploading traffic.X-axis indicates the bit rate the background application generates.

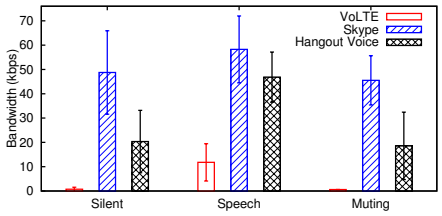

Figure 4: RLC uplink throughput in three scenarios. Silent is when only background noise is presented and muting is when call is intentionally muted.

the radio link control (RLC) layer throughput, which is the aggregation of both audio traffic and control traffic to quantify the total data consumption of different applications.

Result. Fig. 4 shows the throughput comparison results. In normal speech case, VoLTE consumes less network resources by 80% compared with Skype and 75% compared with Hangouts Voice. The difference comes from various factors including some LTE technologies that only available for VoLTE such as discontinuous transmission (DTX) [5] and discontinuous reception (DRX) [15], and the unique codec support [3]. DTX, for example, is a LTE feature, but currently only available for VoLTE, which allows a device to reduce transmission frequency when "no speech" is presented in the conversation. We also observe that AMR-WB codec used by VoLTE only generates minimum Real-time Transport Control Protocol (RTCP) packets to keep link liveness when the call is on hold. These features of VoLTE function together to reduce the transmission rate.

For "Speech" phase, we break down the traffic and examine the throughput of audio packets. In general, Skype generates 3.0× and Hangouts Voice 2.4× audio traffic of VoLTE when transmitting the same audio sample. We also find that VoLTE's rate adaption is more aggressive and efficient in reducing unnecessary transmissions. Compared with the codec adopted by VoIP applications, AMR-WB codec used by VoLTE is more sensitive to the variance of user's voice activity patterns, especially when silence or user-intended muting is presented. In the silence case, AMR-WB is identified to increase its packet generating interval from 20ms to 150ms, which reduces the required network resources significantly.

3.4 Mouth-to-ear Delay and Jitter

As discussed in §2, the difference between VoLTE and OTT VoIP mainly comes from the dedicated bearer. VoLTE is able to use the dedicated bearer that has bit rate guarantee to provide better Quality of Service (QoS), which is 49kbps as observed in our experiment under good signal strength, while the delivery of all other application traffic doesn't have such performance guarantee. Note that ENodeB may decrease the guaranteed bit rate for devices in undesirable network conditions.

Methodology. To validate the QoS guarantee provided by dedicated bearer, we perform VoLTE and OTT VoIP calls under various network conditions and compare the QoS metrics. We use the Spirent Nomad device mentioned in §3.1 to get the accurate *mouth-to-ear delay*, which is defined as the latency between the time when the speaker utters a word and the time when the listener actually hears it [10], containing both the one-way latency in the network and the time spent on encoding and decoding audio packets. We also calculate the packet loss rate and jitter based on the network trace of over 10 hours voice call for each application.

Result. Fig. 5 plots the jitter and mouth-to-ear delay comparison result. VoLTE has 24% less tail jitter and 64% less mean mouth-to-ear delay compared with OTT VoIP. According to the ITU-T standard [29], the 153 ms mean mouth-to-ear delay of VoLTE suggests excellent user satisfactory, while the over 300 ms delay of OTT VoIP indicates some users are not satisfied. In some cases the mouth-to-ear delay of OTT VoIP even exceeds the 400 ms upper bound specified in the standard, which has very negative impact on user experience. Note that the performance gain of VoLTE on the mouth-to-ear delay comes not only from the dedicated bearer, but also from the native codec support in the device firmware [38]. OTT apps adopts user space codec implementation, whereas the embedded VoLTE integrates with low-level audio drivers in the firmware for audio capture and rendering, which is more efficient. In addition, the average packet loss rate of VoLTE among the experiments is 0.009%, which is also clearly better than the 0.03% packet loss rate of Skype and Hangouts Voice.

3.5 User Perceived Call Setup Time

Methodology. We define the *user perceived call setup time* as the time between the caller's click of the "Make phone call" button and the incoming call UI shows on the callee's screen. To support automated measurement of this metric, we use QoE Doctor [18], which integrates UI automation and can be easily adapted to perform automated calls and capture UI updates. We conduct 200 times of call setup experiments for each application in good network condition.

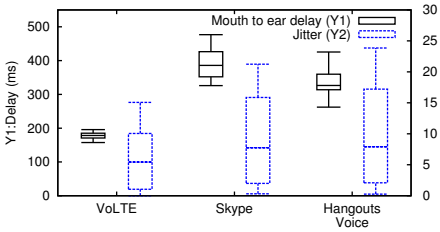

Figure 5: Jitter and mouth-to-ear delay comparison among different applications.

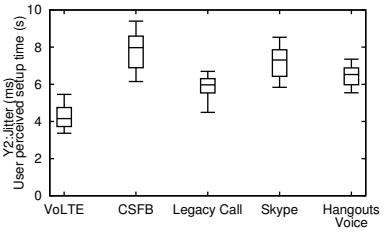

Figure 6: End-to-end call setup time comparison among different applications.

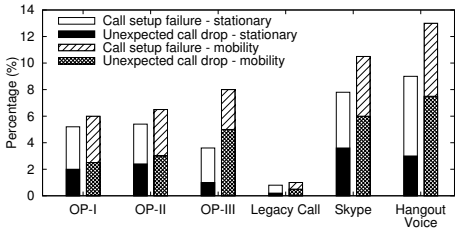

Figure 7: Call reliability comparison among different applications in stationary and mobility experiments.

Description	Mean power (mW)	Std. dev.
Baseline	14.88	134.51
VoLTE	888.74	45.35
Legacy call	511.00	474.98
Skype	2027.06	495.06
Hangouts Voice	2029.53	530.58

Table 3: Power consumption of different applications

Result. Fig. 6 shows that the mean VoLTE call setup time is about 4 s, which is much better than legacy call. In general, VoLTE reduces 48% of user-perceived call setup time compared with CSFB and uses around 40% less time compared to OTT VoIP. The reduced call setup time compared to legacy call and CSFB mainly owes to the simplified IMS based call setup procedure [13].

3.6 Energy Efficiency

VoLTE is expected to have better battery performance compared with OTT VoIP applications when transmitting audio traffic in LTE. According to the technical specifications, the energy benefit mainly comes from these aspects: (1) native codec support reduces the overall CPU computational overhead and power consumption; (2) DTX and DRX reduce the state transitions for transmitting and receiving packets in the silent case; (3) semi-persistent scheduling (SPS)) [32] reduces control channel overhead for VoLTE with radio resource preallocation.

Methodology. We use Monsoon power monitor [26] to calculate the power of different voice call applications. The baseline power is calculated when screen is off and all the background services except mandatory system services are disabled. Each of the result comes from a 10-minute power measurement. Table 3 shows the power consumption comparison result. VoLTE consumes 60% more energy compared with legacy call but uses 75% less energy compared with Skype and OTT VoIP, which conforms with users' expectations.

3.7 Result Summary

Our empirical measurement suggests that VoLTE is a promising technology that integrates the support from both device manufacturer and cellular network providers to deliver excellent audio quality with 83% less bandwidth consumption, around 40% less user-perceived call setup time and 75% less energy consumption compared with OTT VoIP.

4. CALL RELIABILITY STUDY

The ability to make and maintain a phone call is also very critical to good user experience. We conduct both stationary and drive tests to examine and compare the call reliability of all the three operators' VoLTE services, together with OP-I's legacy call and OTT VoIP.

4.1 Call Reliability

We define a "successful call" as one in which the call is successfully established and is maintained for the 1-minute duration of the test, so the call reliability can be measured as the *probability of making a successful call*. We consider two types of failures that lead to an unsuccessful call in our experiment. *Call setup failure* is defined as the call session failing to be established, while *Unintended call drop* is defined as a call drop during an active call session and is not intended by users.

Methodology. For each call application, we perform 500 automatic calls under good signal strength in the stationary reliability test. For mobility case, we conduct drive test on 2 routes, where inter-RAT handover will happen for all the three operators. 100 calls are performed for each call application on each route. We classify the unsuccessful calls according to different failure causes extracted from log analysis and characterize the reliability of different call applications.

Result. Fig. 7 shows the comparison results of call reliability in both stationary and mobility cases. Our experiments show that the reliability of VoLTE and OTT VoIP calls are still not comparable with legacy CS calls, especially in mobility scenario. The root cause is the immature LTE coverage since most of the problems occur when LTE signal strength is low or during the inter-domain switch. Interestingly, we find that the call drop rate of OP-III's VoLTE service significantly increased when moving from stationary test to drive test. We confirm that neither CSFB nor SRVCC are adopted by OP-III, which keeps the simplicity of the LTE network at the risk of a high call drop rate. However, we also notice that even with CSFB and SRVCC adopted, the VoLTE call reliability of OP-I and OP-II is still unsatisfactory. It is crucial to identify the root causes and resolve these reliability problems in the current VoLTE deployment.

4.2 Audio Quality Monitor

However, our call reliability study fails to capture all the VoLTE problems that the end users are experiencing. Some audio experience related problems such as muting and one-way audio [46, 19, 49], cannot be easily detected from log analysis, but have great influence on user experience. To capture such audio experience related problems in real time and understand their impact, we devise a tool called audio quality monitor to detect and report three types of common VoLTE problems with high accuracy.

4.2.1 Audio Problem Detection

Audio quality monitor listens to the phone call state changes and intercepts the incoming audio of a call session to examine the audio quality in real time and report audio problem if detected. Differing from previous work that study performance of real time communication applications using network traffic [24, 25], our approach uses the audio stream output to earphone to study the audio per-

formance. We claim two advantages of our audio stream based analysis. First, network traffic based analysis has accessibility issue on unrooted devices, thus cannot be widely deployed in production network. Even worse, VoLTE audio traffic is transmitted and received in baseband processor [41] and there is no other way to access VoLTE packets than using external tools such as QXDM. Audio stream based analysis overcomes such problem by intercepting the audio channel to get essential information concerning audio quality, which doesn't require root access to the device. Second, audio based approach is closer to real user experience, and can capture some problems that cannot be identified from network trace analysis such as audio decoding failure.

Incoming voice is sampled as byte streams in real time using the Android `AudioRecord` API [11] and we identify audio problems by monitoring the sample buffer. For example, the pattern of a muting problem is defined as consecutive zero buffers presenting in the sampled byte stream and parameters we can configure are the sampling period and the threshold value of reporting a problem. A over-10-second muting is considered intolerable according to the standard [31], we configure the parameters to make the audio quality monitor responsive to muting over 1 second in order to also capture problems such as intermittent audio. This detection algorithm can also distinguish problematic muting from silence, since there are still non-zero bytes received during silent period. To ensure privacy, the audio quality monitor doesn't record any audio and only monitors a sample buffer that flushes frequently.

Using the audio quality monitor, we can capture various telephony audio problems such as muting, intermittent audio and garbled audio besides call setup failure and unintended call drop. We denote these audio related problems as *audio muting* in the rest of this paper, since there is usually short or long periods of muting involved. A context collector is also integrated, which intercepts the `onSignalStrengthChange()` callback function to trace the network environment changes. The collected problematic context information will be uploaded to the server and be analyzed later for diagnosis purpose.

4.2.2 A Useful Diagnostic Tool for Operators

Voice quality measurement through audio stream analysis has been proposed by ITU and has been investigated by some previous work [30, 35, 42]. We present an elegant on-device implementation compared to some existing solutions that are implemented on stand-alone equipment. We claim that the usability of audio quality monitor makes it an efficient diagnostic tool for operators to deploy in their production network. After investigating current problem diagnosis approaches adopted by the three carriers, we find that although some data collection tools are used to help their customer care systems resolve user complaints, the disclosure of many end-user problems still largely relies on user tickets, which are subjective and contain only coarse-grained information. The audio quality monitor overcomes this challenge by providing not only objective problem reports, but also corresponding context information that can assist future diagnosis. Moreover, the capability of reporting problems in the real time also enables remedial actions to be taken by applications to reduce the QoE impairment when encountering problems.

4.2.3 Evaluation

We evaluate the accuracy and overhead of the audio quality monitor in this section. We conduct both a subjective user study and an objective fault injection based evaluation to show the problem detection accuracy. Table 4 shows the summary of the results.

User study evaluation	FP: 0.65%(3/462)	FN: 5%(1/20)
Controlled evaluation	FP: -	FN: 3.7%(37/1000)
Energy consumption	+7% during VoLTE call	
Data usage	1.12 KB on average for each call	

Table 4: Audio quality monitor evaluation result

Tool accuracy: user study. We distribute audio quality monitor to 2 groups of people, ten students from a university and three engineers from one of the operators. With all the participants' consent, we run our app in the background on their personal devices to monitor each call session and upload the results to our server on Google App Engine. For the purpose of evaluation, a feedback window pops up after each call to let the user select whether they have encountered audio muting problem during last call session. Among the 482 calls performed during two weeks, 20 problematic calls are reported by users. Comparing with our detection result, we get the false positive rate as 0.65% (3/462) and the false negative rate as 5% (1/20). After tracking the destination number of the false positive cases, we find that they are all made to some self service numbers, such as airline service, and muting happens when user is interacting with the automatic speech, which will reported by our audio quality monitor, but won't be considered as a problem by users. Since the number of problematic calls is not sufficient to get a convincing false negative rate, we conduct another automated test to evaluate the false negative rate.

Tool accuracy: controlled evaluation. By pressing the "Mute" button during a phone call automatically, we simulate the audio muting problem of adjustable duration to test our detection accuracy. We perform 1000 VoLTE calls with 1-minute call duration and inject 1$\sim$2 seconds muting randomly for each call. Our system is able to detect 96.3% of injected muting issues. After investigation of the false negative cases, we find that the cause is the network type change happening during the muting, which may cause codec changes, or switching between different audio decoder modules. These changes may cause some non-zero bytes appearing in the byte stream during the muting and thus violates our defined pattern of muting problems.

Tool overhead. Comparing the MOS of the VoLTE call with and without audio quality monitor running in the background, we find that audio quality monitor doesn't affect the audio quality. and it only introduces 7% more power consumption during the call session compared with the baseline VoLTE call. Based on our user study data collected for two weeks, the average data consumption of uploading the context information is 1.12 KB for each problematic call.

4.3 Results Summary and Reflection

With the audio quality monitor loaded on the testing devices, we perform a large amount of automated VoLTE call tests in both OP-I and OP-II's network to study VoLTE problems. 3869 VoLTE calls are made in 5 different locations in two cities in US, with durations ranging from 0 to 5 minutes, which is approximately the average wireless call length in US[20]. Fig. 8 shows the occurrence of different types of failures. Note that we also consider failing over to CSFB as a minor problem of VoLTE. CSFB helps reduce VoLTE call setup failure, but will lengthen the call setup time (up to 10 seconds), and may cause temporary out-of-service after the call [52]. The results conform with our expectation that the reliability of a VoLTE call is closely related to the network condition.

We are then motivated to further discover and study these problematic cases under undesirable network conditions. We borrow the idea from "Stress testing" in software engineering, and leverage

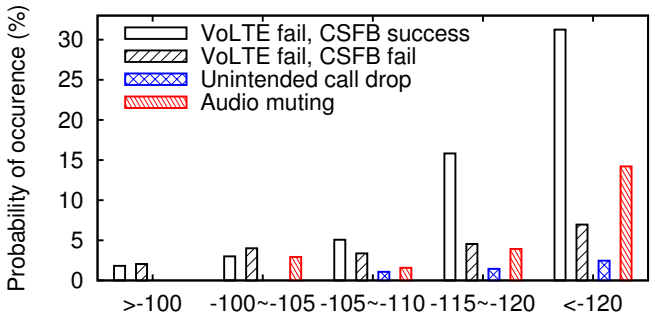

Figure 8: Occurrence of VoLTE problems under different signal strengths. "VoLTE fail, CSFB success" denotes VoLTE fails over to CSFB and establishes the CS call successfully. "VoLTE fail, CSFB fail" denotes the VoLTE call setup failure even with CSFB attempt.

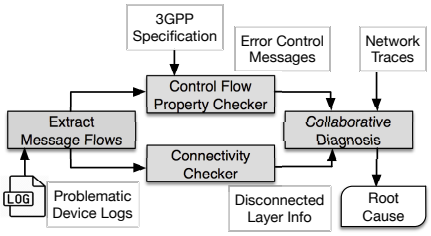

Figure 9: Cross-layer root cause diagnosis flow

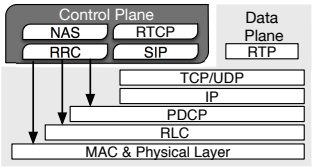

Figure 10: VoLTE control and data plane protocol stack.

fine-grained control of network condition to conduct VoLTE experiments in lab settings. We present our methodology of reproducing and diagnosing these problems in §5.

5. STRESS TESTING AND DIAGNOSIS

We take the stress testing approach to further investigate VoLTE problems in undesirable network conditions. By tuning the network condition worse, we expect more VoLTE problems to be captured by our tool, thus root causes can be identified from the problematic logs. With the help from one of the operators, we gain the control of inter-cell and inter-RAT signal strengths, which enable us to control network events such as inter-cell handover and inter-RAT handover easily in lab settings.

5.1 Stress Testing

The stress testing environment is built upon many equipment that can work together to simulate a given network environment and can give access to different layers' device information, which is hard to obtain before. Besides the previously mentioned RF shield box, antenna and attenuator we used to control the device received signal strength. We further leverage the control of LTE and 4G cells' signal strengths to manipulate various kinds of handover events to simulate the mobility scenario.

We collect QXDM logs, which contain information from chipset, including over-the-air signaling messages and lower layer protocol data units (PDU), `tcpdump` traces, and devices dump state log, which contains SIP messages and debug messages of various system services. These information is not accessible using common methods, but will provide essential information for diagnosing problems on device side and can help troubleshoot problems on network side.

5.2 Cross Layer VoLTE Diagnosis

Previous work [50] proposes a diagnostic approach, that uses model checking to uncover problematic protocol interactions in cellular network. Lesson learned is that problems in cellular network are very likely caused by the interactions between different layers, different entities and different domains. Further motivated by our call reliability measurement and collected problem reports from user study, we take the cross-layer diagnosing approach that investigates both control plane and data plane interactions across multiple layers, and between device and network. We describe the workflow as outlined in Fig. 9.

Extract message flow process takes the massive logs, including QXDM trace and other context information sorted with chronological order as input and output the organized messages flows of each layer. Fig. 10 shows the control plane and data plane protocol stack associated with VoLTE. For control plane traffic, interactions between device and network will be examined in four layers, including RTCP, SIP, Non-Access Stratum (NAS) and Radio Resource Control (RRC). For data plane, flows in another 4 layers, which are RTP, IP, Packet Data Convergence Protocol (PDCP) and Radio Link Control (RLC) will be considered. Data flows and control plane messages will be processed separately later.

Control flow property checker checks the message flows of multiple layers' during a problematic call session, with the intact control flows specified in the 3rd Generation Partnership Project (3GPP) [9] standard. Once violation with specific message in the specification is found, the violation, together with the error messages will be output to the collaborative diagnosis process. For example, network responding with SIP 503 "Server Unavailable" error message to device's SIP 183 "Session Progress" message will be considered as a violation since the SIP 503 error is not supposed to appear after SIP connection has already been established.

Connectivity checker accounts for the data disconnections if occurring during the call session. By calculating the real-time throughput of the four data plane layers, connectivity checker can identify which layer's disconnection causes the audio problem. Observation from the analysis of several instances of problems is that it is not always the case that lower layer's failure causes the upper layer's disconnection, upper layer such as RTP may fail individually due to some control plane problems, thus causing VoLTE service interruption perceived by end users. These layers' connection failures will also be reported to the collaborative diagnosis process.

Collaborative Diagnosis is the core process of the VoLTE diagnosis system. We report the problems to one of the operators, if the problem is not purely caused by device. We then take the corresponding network side logs provided by the operator into analysis, which are captured on base stations, and can be leveraged to verify problems reported from device and diagnose network-originated faults.

5.3 Limitation

Diagnosis process is not fully automated. Among the four processes shown in Fig. 9, checking the control flow property with 3GPP and the collaborative diagnosis still rely heavily on manual inspection. Our current design is the first step towards a highly au-

Symptom	Impact on user experience	Carrier[1]	Potential cause
Unsuccessful call setup	Unable to make any phone call	OP-I OP-II	A lack of coordination in device-network interaction
Unintended call drop	Unable to maintain a phone call	OP-I OP-II	Incorrectly ordered inter-dependent actions
Long Audio muting	Up-to-50-second unrecoverable muting followed with call drop	OP-I OP-II	A lack of coordination in cross-layer interactions

[1] These cases are concerned with system complexity and do not apply to OP-III due to its simpler design choices (e.g., no CSFB and SRVCC support). However, this simplicity leads to other reliability problems as shown in §4.1.

Table 5: Case study overview.

tomated cross-layer diagnostic tool and it will be our future work to build a diagnostic framework to systematically reproduce and diagnose problems.

Checking the control flow property with specifications is non-trivial. A suggested approach is the use of network simulation tools [34, 12]. For example, the Anritsu Signaling Tester [12] implements the protocol stack following the specifications and gives users the control of all the signaling messages. By replaying the traces in the simulated network, violations with the intact flows can be detected automatically.

Discovered root causes may not apply to every instance of problems. Although the symptoms of VoLTE problems we discovered in our stress testing are the same as problems that end users have encountered, there may be various causes rooted in a single symptom. We are working with one of the operators to fix these problems in their network, after which, the impact of the problems we discovered could be quantified.

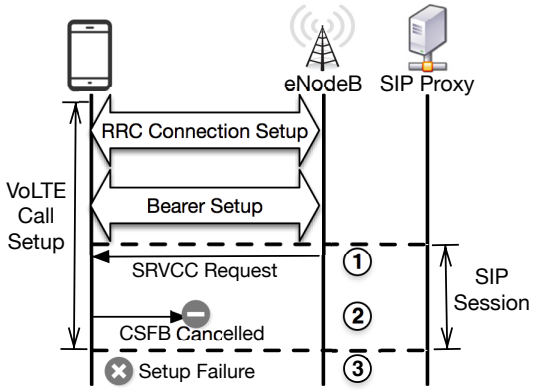

Figure 11: Call setup failure due to untimely SRVCC request.

6. CASE STUDY AND ROOT CAUSES

Table 5 summarizes the uncovered problems in the VoLTE related protocol design and implementation, which contribute to the three most common VoLTE problems affecting users nation wide. Note that the problems related to VoLTE's handoff support don't concern OP-III's VoLTE service since it doesn't have such support. However, simply disabling the handoff support including CSFB and SRVCC is not the panacea, we provide the discussion in §7.

We discuss the uncovered VoLTE problems, deduce the root causes with the help of the diagnosis support, and propose solutions. Since these problems are identified in the telecommunication protocol's design, implementation and configuration, proposed solutions consider all the three entities involved: (1) protocol designers, who develop standard such as RFC and 3GPP; (2) technology vendors, who implement protocols, e.g., Qualcomm and Ericsson; (3) Operators, who configure the products from vendors for deployment.

6.1 A Lack of Coordination in Device-originated and Network-originated Events

Our automated stress testing reports a high call setup failure rate when making VoLTE calls below certain RSRP threshold (-110dBM, 3 out of 5 signal strength bars). From QXDM logs, we observe that VoLTE call setup terminates due to SRVCC failure and CSFB's initiation failure, which is abnormal since CSFB is supposed to redirect the call to CS domain when VoLTE call setup fails.

Problem analysis. Shown in Fig. 11, the call setup failure is caused by problematic interactions between device and network during VoLTE call setup. CSFB is a mobile-originated event and serves as the alternative of VoLTE when LTE condition is not good enough to establish a VoLTE call. We inferred that the trigger for device to originate CSFB is a timer for receiving SIP provisional responses from network during the VoLTE call establishment. Upon expiration, device terminates the VoLTE call setup process and initiates CSFB.

SRVCC, which is a network-originated events, may also be triggered during call setup. Upon receiving measurement report from device, ENodeB may decide to initiate SRVCC if signal strength in a past period of time is considered bad. After the preparation is finished, ENodeB sends SRVCC request to device to handover the ongoing VoLTE session to CS domain (①). Problem occurs when SRVCC request reaches device before CSFB is originated, in which case, device is usually waiting for the SIP timer to expire, in order to initiate CSFB. Device then stops waiting immediately, cancels CSFB process and attempts to do SRVCC since the signaling from network has higher priority (②). However, call will be dropped due to SRVCC failure, since SRVCC works only for established VoLTE sessions, but the call has not yet been set up in this scenario. (③).

Root Cause. From the perspective of protocol design, the cause of such problem is that the specifications [6, 7] fail to coordinate SRVCC and CSFB, which have the same semantics in the call setup context. Since the device-originated CSFB and the network-originated SRVCC both try to switch the call to legacy CS domain during the call setup, it is reasonable to have them coordinated in the specification, so that device can understand the semantics in network-originated request and initiate the handoff process instead of incorrectly releasing the call.

Suggested solutions. We suggest protocol designer to coordinate these two events, which can be done by adding logic to CSFB and SRVCC specification to handle such case that untimely SRVCC request is received during call setup time.

Lessons. Protocol designer should be cautious when designing protocols that involve multiple entities since the interactions among entites may lead to unexpected situations in real time. Adding the

logic to handle unexpected events during protocol design is always recommended.

Follow-up. 3GPP Rel-10 introduces a new type of SRVCC called aSRVCC [2], which supports the SRVCC packet-switched to circuit-switched handover to be initiated during the alerting phase of the call setup. It solves the problem by adding the logic on device side to handle the SRVCC request received during the call setup. OP-I has started to require OEM to integrate aSRVCC support and we are expecting to see it implemented on the upcoming series of devices.

6.2 Incorrectly Ordered Inter-dependent Actions

Unintended call drops frequently occur in our stress testing, when signal strength is tuned down to -120dBm (2 out of 5 signal strength bars), which is a common signal strength level in rural area and indoor environment. In normal cases, SRVCC successfully initiates and the call continues in the CS domain. However, abnormal cases are seen that SRVCC fails to initiate due to the device having handed over to the non-LTE domain. From QXDM log, we confirm that it is because the early arrival of inter-RAT handover request that switches device to legacy network. Without SRVCC initiated, such handover results in call drop.

Problem analysis. In a VoLTE call session, ENodeB monitors the measurement reports from device continuously to decide the triggering of several events. SRVCC and inter-RAT handover are two of them that are of interest in this case. SRVCC is designed to be an improvement on the existing inter-RAT handover in terms of keeping the ongoing call after the handover process. It is apparent that SRVCC should be initiated always before the inter-RAT handover to keep the continuity of VoLTE call in undesired LTE network conditions or moving out of the LTE coverage. However, we find that the sequence between such inter-dependent events is not enforced properly, which causes inter-RAT handover request being delivered to the device before SRVCC request. The late-arriving SRVCC request then fails to be served since device is already redirected to the non-LTE network, which leads to call drop after the handover.

Root cause. The problem is reported to one of the operators and they notice that it can be caused by misconfiguration of the threshold values for SRVCC and inter-RAT handover in the ENodeB. When SRVCC is integrated into the network, existing threshold values should be adjusted to guarantee that inter-RAT handover never occurs before SRVCC. However, even after verifying the proper threshold, call drops due to SRVCC failures are still captured, which indicates that incorrectly ordered delivery of the signaling messages still exists. We identify two causes. First, as shown in Fig. 12, upon SRVCC initiation, core network receives request from ENodeB and initiates call continuity procedure with IMS (①). When inter-RAT handover is triggered, ENodeB only needs to wait for core network to complete handover preparation (②). The first instance of the incorrect ordering appears when the response of SRVCC is delayed due to processing and reaches ENodeB later than the inter-RAT handover response (③). Second, incorrectly ordered delivery can also happen due to the unreliable transmission of the signaling channel between device and ENodeB (④), which has been discovered by previous works [50]. In both cases, VoLTE call will drop after handover since SRVCC fails to be initiated to ensure call continuity.

Suggested solution. Thinking out of the scope of the reliability problem of transmission in cellular network, inter-RAT handover is actually redundant for VoLTE call, since SRVCC inherited and improved all its functionalities in VoLTE scenario. We suggest operators to turn off the inter-RAT handover for all VoLTE call sessions, which can be done by disabling inter-RAT handover for all dedicated bearers on ENodeB.

Lessons. When implementing network protocols, it is not reliable to enforce the inter-dependency implicitly by using some configurable values. For example, using the values of signal strength threshold to ensure sequence of events is shown to be not robust in this case. Explicit enforcement of sequence, for example, specifying that event B can only be trigger after event A has been tried, is a recommended solution.

Follow-up. OP-I has turned off the inter-RAT handover for all dedicated bearers in some markets to evaluate its effectiveness in reducing the VoLTE call drop rate. If it turns out to be effective, this change will be applied in a larger scale.

6.3 A Lack of Coordination in Cross-layer Interactions

We discover extremely long VoLTE audio muting problem in our stress testing, when radio link disconnection is injected to an ongoing VoLTE session. We further find out that a merely 3∼4 seconds' radio link disconnection can lead to an up-to-50-second unrecoverable muting followed by call drop. Such long service interruption time may seriously damage user experience.

Problem analysis. Our cross-layer diagnosis discovers the cause rooted in the data plane protocol interactions. VoLTE audio packets are transmitted based on RTP protocol, where RTP timeout is used to determine the ending of a call session if no packets are received for a period of time. The configuration of the RTP timeout refers to the RTCP interval specified in its RFC specification [43], whose minimum value is recommended as *360/bandwidth(kbps)*. We infer the RTP timeout configured in both OP-I and OP-II's network, which are 30 and 45 seconds. They conform with the recommended value since the bit rate of both OP-I and OP-II's AMR-WB codec are fixed to be 12.65 kbps based on our experiments.

However, we find that this long RTP timeout far exceeds the maximum possible lower layers' recovery time in the context of VoLTE, which leads to unnecessarily long service interruption time when radio link problem happens. Fig. 13 shows the cross-layer interaction when radio link problem occurs. Upon radio link disconnection (①), RLC starts retransmitting unacked packets until a *maximum retransmission threshold (maxRetxThreshold)* is exceeded, and a radio link failure event is reported to RRC layer (②). RRC layer then attempts to reestablish the connection, but will stop trying if certain timers expire, when waiting for the network response to the re-establishment request (③). T301 and T311 denote the timeouts of waiting for two types of responses [8] to the RRC re-establishment request. On expiration, device goes to RRC_IDLE state and the radio link failure will never recover in this session. The aggregated recovery time (T) of layers below RTP can be denoted as:

$$T = (t * maxRetxThreshold) * N + min\{T_{301}, T_{311}\} \quad (1)$$

where t denotes the time interval of transmission and can be approximated by a RLC layer RTT, which can be assumed as 100 ms [40]. N denotes the number of RRC connection re-establishment attempts.

The T_{301} and T_{311} timer are all hundreds of milliseconds inferred from our experiments, which means that an over-one-second radio link disconnection during the RRC re-establishment process will cancel the further attempts of recovery ($N = 1$). We further get the $maxRetxThreshold$ value from QXDM logs that it is configured 16 or 32 by different operators. Then the estimated recovery time of radio layers are all within several seconds for both opera-

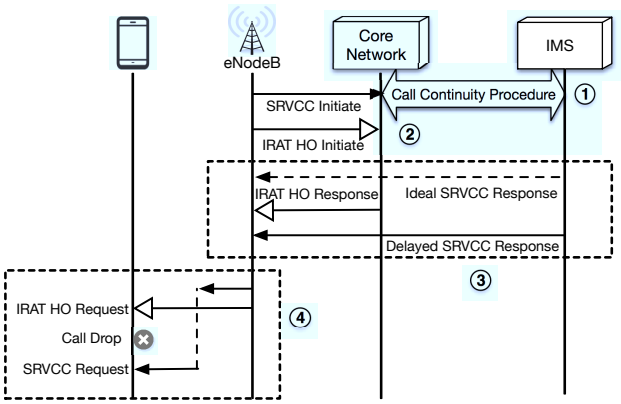

Figure 12: Call drop due to incorrectly ordered messages delivery of SRVCC and inter-RAT handover (IRAT HO).

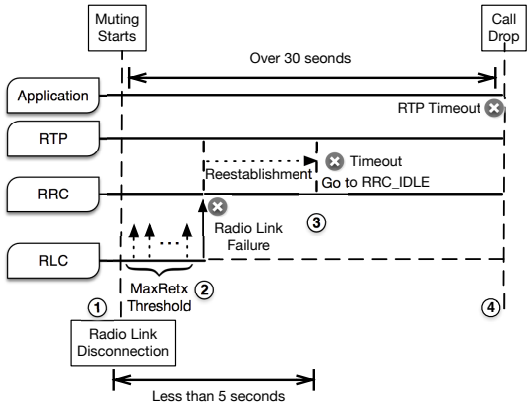

Figure 13: Long audio muting due to lack of coordination among cross-layer timeouts.

tors calculated using the formula (1). The inconsistency between the over-30-second RTP timeout and the several-second recovery time results in the extremely long audio muting.

Root cause. The root cause is identified as the lack of coordination among the design of different specifications. In RTP/RTCP protocol, which is standardized by the Internet Engineering Task Force (IETF), the design principle of the RTCP interval is to limit the portion of control plane traffic over the wire under multiple senders [43]. It doesn't consider the underlying layers' failure recovery mechanism since they are not defined for the use case in the cellular network. The gap between the specifications needs to be bridged if the cross-layer interactions are involved.

This lack of coordination among layers is also considered as an information gap between RTP protocol implemented in the OS and radio layer protocols implemented in the chipset. In the current device implementation of VoLTE stack, radio link events are not passed to the RTP layer, so that upper layer cannot react timely to underlying layers' failures, thus resulting in user experience impairment.

Suggested solution. Since RTP timeout exists both on device and network sides and there are other applications relying on this protocol. It is impractical to simply violate RFC specification to reduce the timeout value. We suggest operators to push OEM to report radio link events such as unrecoverable radio link failure directly to VoLTE call applications, so that reactions, such as sending new service request or dropping the muted call can be taken to reduce service interruption time.

Lessons. When integrating external protocols into an existing protocol stack, it is critical to investigate the difference between the original context and current context before implementing following the specifications. Service providers need to ensure the compatibility of protocols defined for different purposes.

Follow-up. We have worked with OP-I to propose a device-reporting schema to two of the major OEMs in the world, which requires baseband to report radio link information, such as VoLTE signaling message, and RTP statistics to the application in real time during a VoLTE call by broadcasting intents. Currently, parts of the functionalities are implemented on prototype devices and we are expecting the full implementation to evaluate its effectiveness.

7. DISCUSSION

Different choices of implementing VoLTE. As mentioned before, OP-I and OP-II integrate fallback support in their VoLTE service to guarantee a shorter time-to-market, while OP-III didn't launch VoLTE until they thought their LTE network is good enough. In the long-term, the design choice made by OP-III to keep their network simple is preferable. However, it takes high risk of poor service availability during the transition period towards an all-LTE network. For example, the VoLTE call of OP-III is recognized to drop immediately once the LTE RSRP is below a certain threshold, which results in a high call drop rate in mobility scenario as shown in §4.1. Despite the different choices, poor LTE coverage is the major cause behind these problems, so improving their network coverage should be kept as a top priority for operators.

Long-term impact of our work. We admit that the problems we identified are transient, if we assume that all the devices will be in good LTE environment and no fallback support is needed in the future. Since all the three instances of problems are related to the underlying protocols of mobile communication, i.e., 3GPP, we believe that discovering and addressing these problems are important and are complementary to the efforts in improving LTE coverage. Our methodology can also be reused to diagnose problems in other services that support heterogeneous network environment, such as Wi-Fi calling to VoLTE handover.

8. RELATED WORK

Cellular network troubleshooting. There have been quite a few work looking into the protocol design and implementation of cellular network to seek for root causes of end users' problems. Previous work [36, 37, 51] have identified loopholes in the charging system of operators' network, which will cause wrong billing for users' data services. Recent work discover problems lying in the data plane [52] and control plane [50] of cellular network, which causes performance degradation and annoying user experience. Other measurement study (e.g., 4GTest [23] and TailEnder [14]) investigates problems in the interaction between device and network that lead to energy waste. Our work discovers problems in VoLTE, which is a new service not covered by previous studies, but is extremely important since making voice call is the most basic need for mobile users.

Voice call performance analysis. For voice call quality analysis, most of previous studies focus on VoIP service such as Skype. The relationship between Skype's voice streaming efficiency and its forward error correction (FEC) has been measured and tuned for improving user satisfaction [24, 25]. Based on call duration from actual Skype traces, Chen et. al. proposed a model to generate User Satisfaction Index (USI) to quantify the degree of user satisfaction [17]. In comparison, our target is VoLTE, a new carrier-

deployed voice call technology intended for replacing legacy calls, which is more challenging due to higher user expectations and additional complexity in network infrastructure discussed in §2.1.

Recently, some VoLTE performance evaluations are reported from both industry and academia [22, 48], which shares the same goal of characterizing VoLTE performance with us. Compared to them, besides QoS metrics, we take a more systematic way to study various QoE metrics such as user-perceived call setup time and call reliability, which have more direct impact on end users' experience. More importantly, we dig deep into the root causes of newly-discovered VoLTE problems and suggested solutions.

9. CONCLUSION

In this paper, we take the first step to quantify the performance of VoLTE, and study its implication on relevant QoE metrics. We also design and implement a on-device VoLTE problem detection tool to examine the impact of these problems on user experience. We perform stress testing based on insights from in-depth call reliability measurement and uncover 3 major VoLTE problems lying in protocol design and implementation, and suggest potential solutions.

10. ACKNOWLEDGEMENTS

We want to thank the anonymous reviews and shepherd for their valuable comments, and would like to acknowledge the support from Yihua Ethan Guo, Yuru Shao, Pete Myron, Abdulshafil Ahmed, Warren McNeel, Philip Hankins, Matthew Harkins, Toaha Ahmad, and QoE Lab team at T-Mobile. Specially, we would like to acknowledge contributions from Suchit Satpathy from operations team at T-Mobile, who grounded us to the practicality of the problems that customers are facing. This research was supported in part by the National Science Foundation under grants CNS-1059372, CNS-1345226, and CNS-1318306.

11. REFERENCES

[1] Wi-Fi Calling. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/service-provider-wi-fi/white-paper-c11-733136.html.

[2] 3GPP. 3GPP aSRVCC Working Item. http://www.3gpp.org/DynaReport/WiCr-500011.htm.

[3] 3GPP. 3GPP.TS129.238 AMR-WB speech codec. http://www.etsi.org/deliver/etsi_ts/126200_126299/126204/12.00.00_60/ts_126204v120000p.pdf.

[4] 3GPP. 3GPP.TS129.238 TrGW. http://www.etsi.org/deliver/etsi_ts/129200_129299/129238/12.04.00_60/ts_129238v120400p.pdf.

[5] 3GPP. 3GPP.TS146 081 DTX for EFR speech traffic channels. http://www.etsi.org/deliver/etsi_ts/129200_129299/129238/12.04.00_60/ts_129238v120400p.pdf.

[6] 3GPP. 3GPP.TS23.216 Single Radio Voice Call Continuity. http://www.etsi.org/deliver/etsi_ts/123200_123299/123216/12.02.00_60/ts_123216v120200p.pdf.

[7] 3GPP. 3GPP.TS23.272 Circuit Switched CS fallback in Evolved Packet System. http://www.etsi.org/deliver/etsi_ts/123200_123299/123272/12.04.00_60/ts_123272v120400p.pdf.

[8] 3GPP. 3GPP.TS24.331 - 7.3 RRC Timers. http://www.etsi.org/deliver/etsi_ts/146000_146099/146081/12.00.00_60/ts_146081v120000p.pdf.

[9] 3GPP. 3rd Generation Partnership Project. http://www.3gpp.org/.

[10] C. Agastya, D. Mechanic, and N. Kothari. Mouth-to-ear latency in popular voip clients. 2009.

[11] Android. Android audiorecord. http://developer.android.com/reference/android/media/AudioRecord.html.

[12] Anritsu. Anritsu signaling tester md8470a. http://www.anritsu.com/en-US/Products-Solutions/Products/MD8470A.aspx.

[13] G. Association. VoLTE Service Description and Implementation Guidlines. http://www.gsma.com/network2020/wp-content/uploads/2014/05/FCM.01-v1.1.pdf.

[14] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani. Energy consumption in mobile phones: a measurement study and implications for network applications. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 280–293. ACM, 2009.

[15] C. S. Bontu and E. Illidge. Drx mechanism for power saving in lte. *Communications Magazine, IEEE*, 47(6):48–55, 2009.

[16] D. Bubley. The big problem for VoLTE. http://disruptivewireless.blogspot.com/2014/02/the-big-problem-for-volte.html.

[17] K.-T. Chen, C.-Y. Huang, P. Huang, and C.-L. Lei. Quantifying skype user satisfaction. In *ACM SIGCOMM Computer Communication Review*, volume 36, pages 399–410. ACM, 2006.

[18] Q. A. Chen, H. Luo, S. Rosen, Z. M. Mao, K. Iyer, J. Hui, K. Sontineni, and K. Lau. Qoe doctor: Diagnosing mobile app qoe with automated ui control and cross-layer analysis. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, pages 151–164. ACM, 2014.

[19] A. S. Communities. People can't hear me on outgoing. https://discussions.apple.com/thread/6800213.

[20] CTIA. Average local mobile wireless call length in the United States from 1987 to 2012 (in minutes). http://www.statista.com/statistics/185828/average-local-mobile-wireless-call-length-in-the-united-states-since-1987/.

[21] R. Electronics. RF Shielded Testing Test Box. http://www.ramayes.com/RF_Shielded_Test_Enclosures.htm.

[22] S. R. Group. An independent benchmark study of AT&T's VoLTE network. http://www.signalsresearch.com/Docs/LTE%20NA%202014%20VoLTE%20Results%20-%20SRG%20Presentation.pdf.

[23] 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 *Proceedings of*

the 10th international conference on Mobile systems, applications, and services, pages 225–238. ACM, 2012.

[24] T.-Y. Huang, K.-T. Chen, and P. Huang. Tuning skype's redundancy control algorithm for user satisfaction. In *INFOCOM 2009, IEEE*, pages 1179–1187. IEEE, 2009.

[25] T.-Y. Huang, P. Huang, K.-T. Chen, and P.-J. Wang. Could skype be more satisfying? a qoe-centric study of the fec mechanism in an internet-scale voip system. *Network, IEEE*, 24(2):42–48, 2010.

[26] M. S. Inc. Monsoon Power Monitor. https://www.msoon.com/LabEquipment/PowerMonitor/.

[27] J. Industries. JFW Programmable Attenuators. http://www.jfwindustries.com/catalog/Programmable_Attenuators-2-1.html.

[28] ITU-T. Itu-t mean opinion score (mos) terminology. ITU-T P.800.1, 2006.

[29] R. ITU-T and I. Recommend. G. 114. *One-way transmission time*, 18, 2000.

[30] R. ITU-T and I. Recommend. G.720.1 : Generic sound activity detector. 2010.

[31] T. S. ITU-T. E. 855. *Connection integrity objective for the international telephone service*, 1988.

[32] D. Jiang, H. Wang, E. Malkamaki, and E. Tuomaala. Principle and performance of semi-persistent scheduling for voip in lte system. In *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on*, pages 2861–2864. IEEE, 2007.

[33] J. Kallio, T. Jalkanen, and J. T. Penttinen. Voice over lte. *The LTE/SAE Deployment Handbook*, pages 157–187, 2012.

[34] LENA. Lte-epc network simulator. http://networks.cttc.es/mobile-networks/software-tools/lena/.

[35] L. Malfait, J. Berger, and M. Kastner. P. 563&# 8212; the itu-t standard for single-ended speech quality assessment. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(6):1924–1934, 2006.

[36] C. Peng, C.-y. Li, G.-H. Tu, S. Lu, and L. Zhang. Mobile data charging: new attacks and countermeasures. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 195–204. ACM, 2012.

[37] C. Peng, G.-h. Tu, C.-y. Li, and S. Lu. Can we pay for what we get in 3g data access? In *Proceedings of the 18th annual international conference on Mobile computing and networking*, pages 113–124. ACM, 2012.

[38] Qualcomm. Qualcomm Announces First Large-Scale Commercial VoLTE Launch in Japan. https://www.qualcomm.com/news/releases/2014/07/08/qualcomm-announces-first-large-scale-commercial-volte-launch-japan.

[39] Qualcomm. Qualcomm eXtensible Diagnostic Monitor. https://www.qualcomm.com/media/documents/files/qxdm-professional-qualcomm-extensible-diagnostic-monitor.pdf.

[40] Qualcomm. Performance evaluation of adaptive rlc pdu size in hspa+ networks. https://www.qualcomm.com/media/documents/files/qualcomm-research-performance-evaluation-of-adaptive-rlc-pdu-size-in-hspa-networks.pdf, 2012.

[41] Qualcomm. Qualcomm chipset powers first successful volte call with srvcc. https://www.qualcomm.com/news/releases/2012/02/02/qualcomm-chipset-powers-first-successful-voip-over-lte-call-single-radio, 2012.

[42] Qualcomm. Spirent audio and vedio testing solution. http://www.spirent.com/Products/ProLab/ProLab-Video-Quality, 2015.

[43] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson. Rfc 3550. *RTP: a transport protocol for real-time applications*, 7, 2003.

[44] J. Soininen. Transition scenarios for 3gpp networks. RFC 3574, 2003.

[45] Spirent. Spirent Nomad HD Voice Measurement System. http://www.spirent.com/About-Us/News_Room/Press-Releases/2013/05-21-13_Nomad-HD.

[46] T-Mobile. VoLTE One way or no audio: Samsung Galaxy Note 3. https://support.t-mobile.com/docs/DOC-11793.

[47] I. T-Mobile. Diagnostics metrics collection. https://support.t-mobile.com/docs/DOC-2929, 2011.

[48] M. R. Tabany and C. G. Guy. An End-to-End QoS Performance Evaluation of VoLTE in 4G E-UTRAN-based Wireless Networks. In *the 10th International Conference on Wireless and Mobile Communications*, 2014.

[49] Tektronix. VoLTE Troubleshooting. http://www.tekcomms.com/sites/tekcomms.com/files/VoLTE_Troubleshooting_CCW-30697-2.pdf.

[50] G.-H. Tu, Y. Li, C. Peng, C.-Y. Li, H. Wang, and S. Lu. Control-plane protocol interactions in cellular networks. In *Proceedings of the 2014 ACM conference on SIGCOMM*, pages 223–234. ACM, 2014.

[51] 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 *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pages 305–318. ACM, 2013.

[52] G.-H. Tu, C. Peng, H. Wang, C.-Y. Li, and S. Lu. How voice calls affect data in operational lte networks. In *Proceedings of the 19th annual international conference on Mobile computing & networking*, pages 87–98. ACM, 2013.