State Transition Latency Reduction Scheme in the LTE/LTE-A Radio Access Network

Jinseong Lee[1], Jaiyong Lee[2]
Department of Electrical & Electronic Engineering,
Yonsei University,
Seoul, Korea
Emails: [1]jinseong.lee@yonsei.ac.kr, [2]jyl@yonsei.ac.kr

***Abstract*— With its higher bandwidth and lower latency, the LTE/LTE-A network has been successively deployed worldwide, leading to many changes in peoples' lifestyles over the years. Recently, as new demands on the cellular network have appeared from emerging technologies such as the Internet of Things, latency reduction has become a key challenge to supporting emerging technologies and satisfying user expectations. Although latency reduction in the cellular network is a traditional issue that has been studied in various respects, the latency of the LTE/LTE-A radio access network has been undervalued and investigated by very few studies. This article presents the latency issues in the LTE/LTE-A radio access network based on the radio resource control (RRC). Furthermore, we propose new approaches to mitigate involuntary delay by optimizing the procedures of the RRC state transition.**

Keywords*— *Latency reduction; state transition delay; LTE

I. INTRODUCTION

In recent years, new demands for machine type communication (MTC), public safety, sports events, and real-time mobile gaming are arising in mobile networks on the communication platform of the Internet of Things (IoT). One of the differences between the new services based on these novel technologies and traditional applications is the traffic characteristic, which is usually very small but may require lower latency, depending on the application [1]. Thus, to satisfy this emerging demand and expand the business area of the LTE/LTE-A network, latency reduction is a critical issue for the next evolution of communications.

In cellular networks, latency reduction has been studied typically in four aspects: network deployment, end node protocol optimization, middle-box buffering, and radio access network (RAN). In terms of network deployment, network technologies such as a contents delivery network (CDN), software-defined network (SDN), and HTTP cache server can reduce the packet travel distances geographically and prevent bottlenecks in the routing paths. Performance is improved dramatically, and faster responses and loading times are provided for overseas contents. From the perspective of the end-node, procedure parallelization and cross-layer optimization techniques have been studied, such as TCP acceleration, transport layer security (TLS), handshake optimization, and so on. Queuing delay reduction in the middle box has been studied for a long time, but recently in cellular networks, the buffer-bloat [2] problem was newly introduced because memory is cheaper and bigger buffers are used. This problem causes the TCP congestion control algorithm to malfunction, which increases TCP session latency. To resolve this issue, techniques such as adaptive queue size control and rate-based transmission size control are being studied.

Despite advances in these aspects, there has been no major progress in the LTE/LTE-A radio access network since 2008, at the first release of the LTE specification. Latency reduction in a radio access network not only decreases the packet transmission time, but also improves the radio resource efficiency and reduces the buffer requirements in mobile devices. Additionally, as new technologies such as M2M (mobile to mobile) and V2X (vehicle to everything) emerge, that enable direct communication, latency reduction in the radio access network becomes an essential requirement.

In this article, we exploit the latency in the LTE/LTE-A radio access network and introduce the state transition delay issues for packet transmission. We first analyze the RRC state transition procedure and then describe the impact of the RRC state transition delay. Finally we propose methods to alleviate the delay and provide our conclusions for this article.

II. BACKGROUND

A. RRC State Transition Overview

This section presents the delay and a procedure to establish radio resources, and describes the delay impact on the commercial network and well as schemes to mitigate the problem. As shown in Figure 1(a), in the LTE/LTE-A network, there are two different states in RRC: RRC_Idle and RRC_Connected, related with radio resource establishment after device power-on and registration [3]. When user equipment (UE, terminal) or an evolved Node-B (eNB, base station) has a packet to send/receive, the RRC state is changed from RRC_Idle to RRC_Connected for connection establishment. RRC_Connected is maintained as long as there are ongoing packets. eNB keeps monitoring the packet activity during RRC_Connected, and if there is no activity for a specific time period (called the RRC inactivity time), then eNB commands the UE to release radio resources and transition to RRC_Idle to withdraw unused radio resources and save the terminal's power. Typically, five or ten seconds are used for RRC inactivity timeout, depending on the network operator's policy. If this value is too large, more of the UE's power is

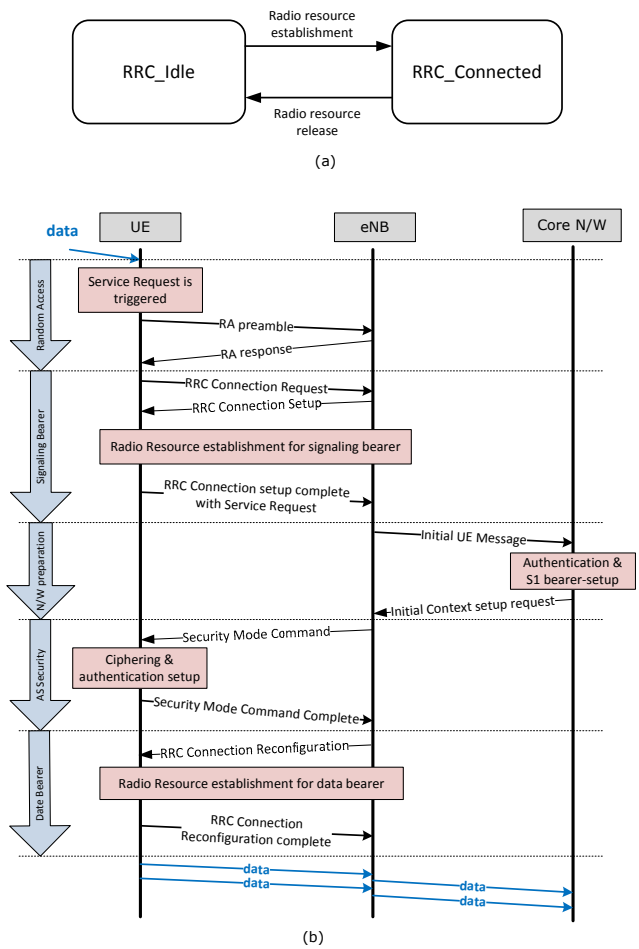

Figure 1 RRC state transition: a) RRC states; b) signaling flow of RRC state transition for data transmission.

consumed and the dedicated radio resource could be wasted. But too small a value may cause excessive signaling overhead and random access collisions due to frequent state transitions. [4]

B. *State Transition Procedure*

The typical call flow for the RRC state transition from RRC_Idle to RRC_Connected is presented in Figure 1(b). It is assumed that the UE is registered and a service request is triggered in the UE for packet transmission. (A network-triggered service request procedure is very similar, except for the paging procedure.) An RRC state transition for data transmission is composed of five parts: random access, radio resource establishment for the signaling radio bearer (SRB), network preparation including authentication and S1 bearer establishment, the access stratum (AS) security procedure, and radio resource establishment for the data radio bearer (DRB).

A new packet received from a higher layer in the RRC_Idle state triggers a Service Request message of a non-access stratum (NAS) to request radio and network path establishment for data transmission. But, because the UE has no radio resources to send this message, through the random access procedures, uplink radio resources are only allocated to send the initial uplink message RRC Connection Request.

Dedicated radio resources for the SRB are allocated by the RRC Connection Request, Setup, and Complete procedures. The Service Request message can be piggybacked and transmitted with the RRC Connection Setup Complete message. This message includes the cause of the Service Request and is bypassed to the mobility management entity (MME). Then, the MME checks the cause and establishes the S1 bearer that provides a path between eNB and the serving gateway (S-GW). After network preparation, the eNB performs the AS security procedures, such as ciphering and authentication. Security algorithms are involved in the Security Mode Command, and all RRC messages are encrypted and checked for integrity protection after a successful AS security procedure.

Finally, the data radio bearer establishment procedure is performed through the RRC Connection Reconfiguration and Complete messages. A reconfiguration message includes the user plane bearer setup parameters for DRB and SRB2. From this moment, the packet that triggers the RRC state transition can be delivered to the network. In 3GPP [5], the control plane delay for the RRC state transition delay is described; it includes random access, SRB establishment, and network preparation, and states that the delay should be less than 100 ms. But when the data transmission delay caused by the RRC state transition is considered, AS security procedure and the data radio bearer establishment delay must also be taken into account.

III. THE IMPACT OF RRC STATE TRANSITION DELAY

In this study, we performed a field test to check the impact of the RRC state transition delay in a commercial environment. The network in the test was KT (Korea Telecom), a major network operator in Korea, and we used LTE Band 3 with a 20 MHz bandwidth. The UE is an LG Nexus 5, and the kernel was modified to not transmit unwilling packets from the background application. The transport layer log was captured using Tcpdump to check the packet transmission time and session connection status. In addition, the LTE modem log was used to confirm the RRC state transition and check the packet delay. This log was gathered from the UE hidden mode and analyzed using the Qualcomm log analysis tool.

Figure 2(a) illustrates the ping round-trip time (RTT) results according to various ping intervals for the well-known websites, Google, Facebook, YouTube, and Twitter. The RTT

TABLE I. ENVIRONMENT SETUP FOR EXPERIMENT

Item	Description
Mobile device	LG Nexus 5 (D810) - CPU 2.3GHz, 2GB RAM - SW version: KOT49H - Android 4.4.2 KitKat (modified)
Network	KT(Korea Telecom)
LTE Band	Band3 with 20MHz
Average bandwidth	13 Mbps
Signal strength	LTE RSRP -93~-99 dBm
Log capture	Tcdump Qualcomm log analysis tool

results were very similar with ping interval under 10 seconds in all cases, but the RTT increased by 120–130 ms with ping interval over 10 seconds for all websites. We found that the RRC state changed from RRC_Connected to RRC_Idle after 10 seconds in the LTE modem log, and that caused an additional RRC state transition delay of 120–130 ms to transmit the ping request. This result shows that the network operator uses 10 seconds as the RRC inactivity timer, and further delay could be generated when the network access interval is more than 10 seconds.

To determine how much this transition delay affects the user experience (UX), we investigated mobile web browsing, which is one of the most commonly used applications in smart phones. The time gap between each web page access is called the mobile web page access interval, and we measured this statistic under the same environmental conditions to determine how long of an interval users use in trying to connect to other web pages. We also captured the test log from Tcpdump and the modem side, and analyzed over 20,000 page views.

The web page access interval is defined as the time between the last TCP session transmission and the new HTTP request during mobile web browsing. As shown in Figure 2(b), 66% of web page view intervals are under 10 seconds, but the other third of view intervals are longer than 10 seconds, meaning these users experience uncomfortable latency. Google reports that every 20 ms reduction in round-trip latency yields a 7–15% reduction in web page loading time [6], and Amazon found that every 100 ms of latency cost them 1% in sales [7]. Considering MTC, V2X, and online gaming, an additional 120 ms of delay may restrict LTE deployment for an industry that has a latency deadline.

IV. Ideas to Mitigate the Impact: Fast RRC State Transition to RRC_Connected

The proposed idea to mitigate the RRC state transition impact is based on the fact that most UE-triggered service requests are the same as previous requests, for some time. If the UE can identify that a request has the same parameters as previous requests and can communicate this information to the eNB, then the establishment procedures can be performed in parallel (i.e., at the same time), and the RRC state transition delay for data transmission can be shortened.

Figure 3(a) shows the RRC states for the proposed scheme, which splits the RRC_Idle state into two different states, S1-Idle and S2-Idle. Hash-Key, and T2 timers are also newly introduced into this mechanism, with the T1 timer being the same as the existing RRC inactivity timer. In S1-Idle (Stage 1), radio resources are released, but RRC keeps the previous configuration and hash-key, which is shared before the radio resource release for the T2 timer. When the UE tries to send an uplink packet in S1-Idle, the UE checks the sameness of the parameters and transmits an RRC Connection Request message with the shared hash-key. If Timer T2 expires without any packet transmission, RRC changes to the S2-Idle state, and the next procedure is identical to the legacy procedure. The detailed signaling flow is shown in Figure 3(b), and the

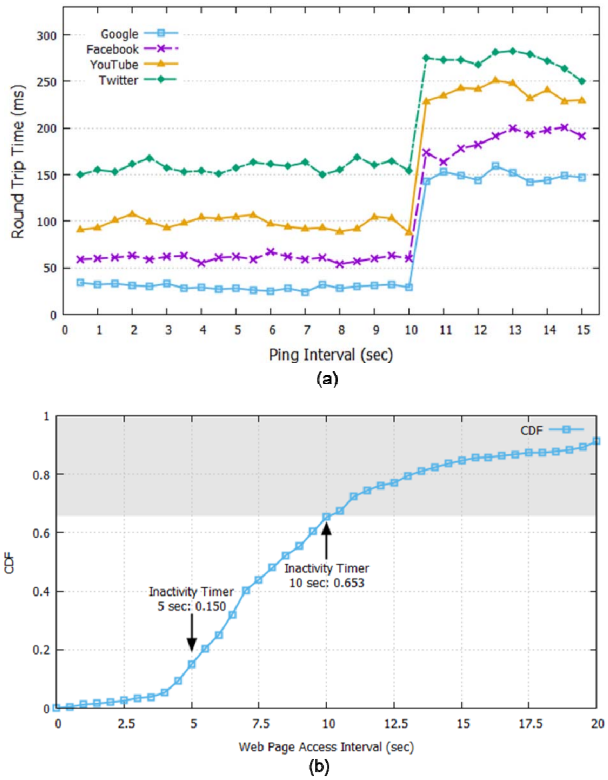

Figure 3 The impact of RRC state transition delay: a) Ping RTT according to the ping interval; b) CDF of web page view interval

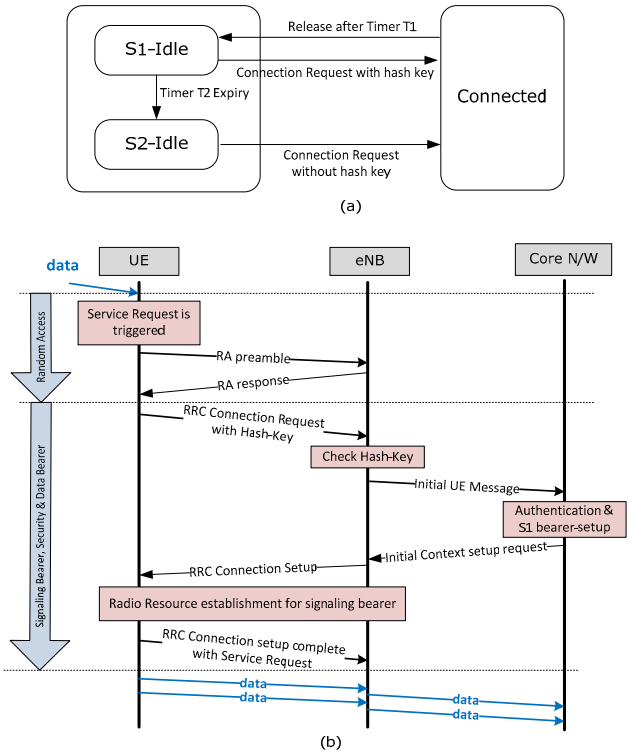

Figure 2 Fast RRC state transition: a) RRC states; b) signaling flow of fast RRC state transition for data transmission

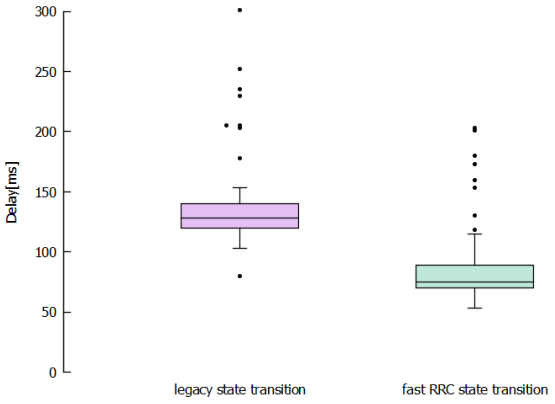

Figure 4 delay comparison between legacy and fast state transition

parameters in the AS Security Mode Command and the RRC Connection Reconfiguration for the DRB can be included in the RRC Connection Setup message. The hash-key may be mismatched between the eNB and the UE under the condition that the cell change during the T2 Timer and the minor timer do not match. In that case, the eNB and the UE each follow the same original service request procedure, and backward compatibility is guaranteed.

V. Performance Evaluation

To evaluate the performance of proposed algorithm in the commercial network, we have captured UE side log and analyzed the delay between each RRC message. Because we cannot change commercial protocol software of UE/eNB, we are considered to be the same the performance of proposed algorithm and the delay due to AS security and reconfiguration procedure. Figure 4 shows that this scheme can shorten the delay caused by the AS security and the data bearer setup procedure and decrease the RRC state transition delay by about 50–60 ms in our analysis, depending on the network architecture.

It is also backward compatible with the previous LTE mechanism without additional signaling overhead. To adapt the idea to a commercial network, the T2 timer could be set to a variable value that depends on the device type. For example, for human-held devices like smart phones, this value would be set to less than 100 seconds, depending on the trade-off policy between memory and complexity. But in the case of public safety devices, which require very low power consumption and as short a latency as possible, Timer T2 should be longer, and the proposed scheme can improve the probability of meeting a latency deadline without any weak point in power consumption.

VI. Conclusion

The traffic characteristics of emerging technologies are different than traditional applications, and as the latency deadline is restricted, shorter and faster response is required. Being an aspect of not only network deployment, middle-box queues, and end nodes, but also radio access networks, latency reduction is a critical issue for LTE evolution. In this article, we introduced the state transition delay in a radio access network and reviewed the impact of delay on applications close to the user, such as mobile web browsing and file transmission. We investigated the delay issue of the RRC state transition, which has the effect on a mobile device of no transmission after inactive timeout. We performed the ping test to measure the inactivity timer in the commercial network and analyzed the web page access interval statistics to confirm the delay effect on mobile web browsing. We then proposed a method to alleviate the problem

References

[1] Ghavimi, Fayezeh, and Hsiao-Hwa Chen. "M2M Communications in 3GPP LTE/LTE-A Networks: Architectures, Service Requirements, Challenges and Applications." (2014).

[2] Gettys, Jim, and Kathleen Nichols. "Bufferbloat: dark buffers in the internet." Communications of the ACM 55.1 (2012): 57-65.

[3] "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol Specification," 3GPP TS 36.331, v12.7.0, Sept. 2015.

[4] Choi, Yongmin, et al. "The impact of application signaling traffic on public land mobile networks." Communications Magazine, IEEE 52.1 (2014): 166-172.

[5] "Technical Specification Group Radio Access Network; Feasibility Study for Further Advancements for E-UTRA (LTE-Advanced)" 3GPP TR 36.912 v12.0.0, Sept. 2014.

[6] Google Blog, "More Bandwidth Doesn't Matter (much)" , April 2010

[7] Liddle, Jim. "Amazon found every 100ms of latency cost them 1% in sales." The GigaSpaces 27 (2008).