Latency in LTE/LTE-A and 5G

Markus Paschke
Ilmenau University of Technology
Integrated Communication-Systems
Email: Markus.Paschke@tu-ilmenau.de

***Abstract*—Latency is playing a tremendous role in LTE-A and upcoming 5G Networks, due to the fact, that new application areas for mobile communications are evolving. Firstly we classify Latency and Reliability Constraints for mobile communications, considering two basic types of networks, namely Long-Term-Evolution-Advanced (LTE-A) cellular networks and medium-range widely deployable mesh networks. Secondly we analyze for the greater part of this paper the medium range networks as an underlaying device-to-device (D2D) infrastructure to establish a working environment for vehicle-to-vehicle (V2V) communications. Finally we will propose approaches and solutions in order to reduce latency and outline further improvements for 5th generation cellular networks.**

I. Introduction

A. Motivation

Machine-to-machine (M2M) or device-to-device (D2D) networks as we call it in this paper have drawn great interest for stakeholders and companies of various sectors like transportation, health care, smart homes, security and so forth. To fulfill the premises every application area is outlining, we have to talk about general requirements such as latency and reliability for cellular and D2D networks. The terms delay and latency are often used interchangeably; we use latency in this discussion as the time to send a short IP-Packet from an User Equipment (UE) to an E-UTRAN Node B (eNB)/Server in the Internet or to another UE, which would be necessary for a D2D Network, and receive a reply back, taking structural additional delay (e.g. encoding at transmitter and detection/decoding at receiver) into account. This latency is mainly caused by the travel time of the signal and the processing of the data at every node it passes through.

In developing wireless 5G standards, we have an opportunity to further reduce latency and time delays to be implemented in future wireless networks. In fact, there appears to be unanimous opinion that 5G standards should have less than 1 millisecond (ms) of latency [1]. Because of the precedent growth in the autonomous vehicle technology it is inevitable to reduce latency to a minimum for fast communication among multiple vehicles to guarantee safety. In a V2V Environment vehicles have a close proximity to each other which require real-time monitoring and therefor strict requirements on reliability and access availability. The EU-Project METIS states, that a maximum delay of 5ms and transmission reliability of 99,999% should be guaranteed [2].

As the minimum possible round trip time of 5ms for a transmission acknowledged by an eNB in the current LTE Networks tested, the actual ping time between a UE and a Server located <50km away from the eNB is largely higher (56ms) [3]. This is due to the fact, that the Core-Network is explicit slower than the direct connection between a UE and an eNB. For the sake of brevity we will concentrate on the approaches to reduce the latency between UE's and UE's or UE's and eNB.

B. State of the Art

Current legacy solutions for V2V communications are adhoc communications over the 802.11p standard and backend based communications over the Long Term Evolution (LTE) cellular standard [4]. Using the 802.11p standard as an underlaying D2D network, which is optimized for a WLAN-Environment with very low mobility. This is leading to unacceptable reliability for fast moving cars or any D2D network involving moving UE's. Additionally, using a current LTE network for backend based communications is not convincing enough regarding latency and reliability for any V2V communication in order to guarantee safety for autonomous cars, traffic control so forth [5]. Using different standards and different approaches for the D2D networks has to be found. A normal scenario using two cars, now stated as two UE's need to communicate directly to each other, sharing the same resources a regular cellular UE would use. This approach would give us four promising gains (i.e. proximity gain, hop gain, reuse gain and pairing gain). Proximity gain enables us high data rates, low delay and low power consumption, because the data would only be transmitted directly to the next car if possible. Skipping the eNB, which would need an up- and downlink, for a transmission is reducing the amount of hops from two to one. Reuse gain implies that the UE's use the same radio resource for cellular and D2D communications, which we stated before. Last but not least, pairing gain leaves the opportunity for UE's to rather connect over an eNB or use the direct connection to another UE for communication mode, depending mostly on resource allocation and interference levels on the network [6].

II. Classification of Latency

The LTE network in general is defined by two parts: The core network (CN) and the radio-access network (RAN). Classifying latency will result in the analyzation of both parts in order to get an overview what causes the most delay in the System. The System Architecture Evolution (SAE) is the core network architecture of the LTE network defined by the 3rd Generation Partnership Project (3GPP). It is a major improvement in terms of latency compared to the 3rd generation mobile networks. Considering a network with less core network nodes will result in multiple benefits:

- Less protocol related processing
- Less number of interfaces

- Minimized cost of interoperability testing
- Easier radio interface protocol optimization (e.g. merging protocol plane protocols)
- Shorter signaling sequences result in more rapid session setup

A. Control-plane latency

The radio protocol architecture is split into the control plane architecture and user plane architecture leaving data flows and controlling messages for the UE and CN seperated, allowing independent scaling for network providers. To limit battery consumption on mobile devices Radio Ressource Control (RRC) timers are set, to go in idle mode (no connection) if the UE did not receive any messages lately. In idle mode it listens to paging messages to know about incoming calls, system information change and Earthquake and Tsunami Warning Service (ETWS) notifications for ETWS capable UE's. Monitoring the physical downlink control channel (PDCCH) continously drains battery power dramastically. So most of the UE's listen at a given period of time for paging messages, which is called discontinous reception (DRX). Going from LTE_idle to LTE_active state however take a small amount of time, because controlling messages have to be sent to synchronize the UE and CN. The delay for transfering a RRC Connection request message with Session Initiation from the UE to the access Gateway (aGW) considering a zero delay over the S1 control-plane is assumed and shown on Table 1. For the complete session setup from idle to active state including the user equipment and core network is a time interval 31ms delay presumed. Detailed calculations are shown by the authors of [7]. After this procedure eNB can schedule the UE for any downlink and uplink transmissions. If the UE has lost uplink synchronization while in DRX mode, it has to use the Random Access Channel (RACH) to receive a timing advance message along with resources for uplink transmission, which leads to another 12-13 ms [7]. In a context of V2V network, cars can be considered UE's and will not have problems with battery drain to the same extend as mobile devices. As a consequence the UE will not enter DRX cycles or LTE_idle states except when being not in LTE coverage and therefor loosing complete network connection. Control-plane latency will affect UE's in a V2V network only for the first and initial setup when connecting to a LTE - Network and therefor neglected in this paper.

UE - eNB		**eNB - aGW**	
UTX L3/L2 processing	1ms	TX processing	0,5 ms
L1 frame alignment	0,25 ms	Frame transmission	0,5 ms
TX L1 processing	2 x 0,5 = 1 ms	RCV processing	1,0 ms
Frame transmission	0,5 ms	**TOTAL**	**2,0 ms**
HARQ retransmissions	5 x 0,5 x 0,3 = 0.75 ms		
RCV L1 processing	2 x 0,5 = 1 ms		
RCV L3/L2 processing	1 ms		
TOTAL	**5,5 ms**		

TABLE I: control plane latency estimation for lte_idle to lte_active transition [7]

B. User-plane latency

User-plane latency is defined in terms of the one-way transit time between a packet being available at the IP layer in either UE or RAN edge node, and the availability of this packet at IP layer in the RAN edge node/UE. The RAN edge node is the node providing the RAN interface towards the core network [7]. The user-plane protocol stack is located between the eNB and UE and consists of the following sub-layers [8]:

- PDCP (Packet Data Convergence Protocol)
- RLC (radio Link Control)
- Medium Access Control (MAC)

To analyze the latency in the user-plane we assume an unloaded network with a single user and therefor no scheduling delays by the eNB as well as no radio interference not causing any hybrid automatic repeat requests (HARQ) retransmissions. Using a small IP-Packet that fits in one subframe with two slots, which is 1ms long and self decodable, will suit our needs because no segmentation is needed. Figure 1 shows that an overall of 3.5 ms one way delay is assumed, which is distributed among the UE, eNB and aGW and will now be identified [7]:

- UE processing delay: header compression, ciphering and RLC/MAC processing
- Resource allocation and physical layer transmission delay: Tx L1 processing, TTI, subframe alignment and Rx L1 processing
- HARQ retransmission delay
- eNB processing delay: RLC/MAC processing
- eNB-aGW delay (on S1 interface)
- aGW processing delay: header decompression and ciphering

The previous assumptions are based on minimum delay conditions. The delay between the UE and eNB are stated as 2ms delay with a transmission time interval (TTI) of 0.5ms over the radio interface. The propagation delay over the S1 interface between the eNB and aGW is stated with 1ms and could travel 200km if it does not travel through any additional nodes. If the packet has to travel to a specific server located in close approximation to the edge node of the core network another delay has to be considered increasing the system

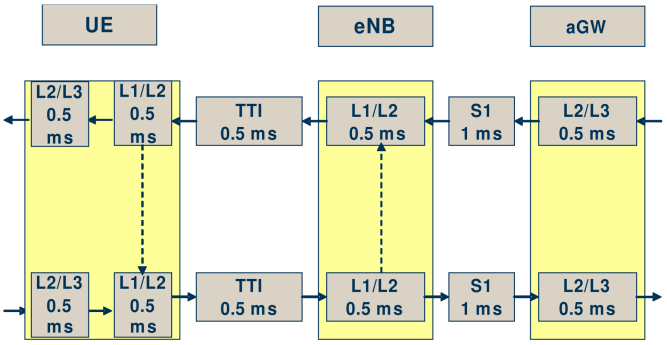

Fig. 1. User-plane latency components [7]

latency. This calculations lead to the a final latency of under 5ms for this specific case of an small IP-packet with no or little scheduling and processing delays, enabling a round trip time (RTT) of under 10ms.

III. Design and Implementation of D2D Networks

In contrast to cellular networks, data is not transmitted via base stations (eNB), which would limit the transmission rate when devices are close to each other, instead a direct link between the UE's is used. As we stated before, this brings benefits such as hop gain, spectrum efficiency, less power consumption and pairing gain. Based on the type of radio-resource spectrum sharing, D2D can be divided into two topology choices: in-band and out-of-band. Out-of-band is utilzing bands other than the cellular spectrum (e.g. 2,4 gHZ bands such as Wifi - Direct / 802.11). In-band can be further classified into two sub groups: An overlaying network which means, that cellular and D2D users use orthogonal time/frequency resources and cellular UE's as well as D2D UE's would not face any interference problems. This approach however is not very cellular spectrum efficient and will not be further discussed in this paper. Enabling D2D links in an underlaying mesh network brings a challenge to radio-resource management. The most critical challenge is the interference between the primary cellular network and the D2D underlay. To handle this resource allocation, a fitting Radio Resource Management (RRM) Strategy has to be found, which includes power control of each UE both in cellular and D2D networks, as well as choosing the right resources to share between them [3]. Finally maximizing the network throughput while guaranteeing Quality of Service (QoS) requirements to D2D and cellular users is essential. The work in [4] presented a RRM strategy as well as requirements for vehicle users (VU's) and cellular users (CU's).

A. Requirements for (VU's) and (CU's)

For V2V networks are latency and reliability constraints one of the most important design aspects. For autonomous driving cars and safety applications, small payload has to be transmitted repeatedly at given intervals, making underlaying networks uplink sensitive. High data rates on the other hand are less important and only for entertainment purposes regarded as a side note.

In contrast to the requirements for VU's, latency and reliability constraints are less strict and a strategy for maximizing the throughput under certain fairness conditions should be considered. With regards to fairness, here we assume the proportional bandwith fairness [9] among CU's that means the number of Ressource Blocks (RB's) allocated to the $m'th$ CU $E'_{m'}$, during one scheduling time unit is given for all $m' \epsilon M'$ and $\sum_{m'=1}^{M'} E'_{m'} = F$ [4].

B. System Model

Considering an approach the authors in [4] stated CU's and VU's share the same uplink radio resources inorder to improve spectrum efficiency and the underlaying D2D network is only used by vehicle users. A D2D communication has mainly two phases: discovery phase, in which devices know their excact location and detect surrounding devices/services and communication phase in which each device is able to transmit data on a D2D link to another device. Cause of fast changing links, caused by fast traveling cars, a lightweight D2D direct discovery mechanism is chosen. To ensure security for D2D users, this session setup is assisted by the CN and optimized for QoS-enabled D2D communications through the eNB, minimizing the involvement of EPC entities (MME and HSS) in order to avoid overhead. After the session initiation between two D2D links, a direct communication path is available and can be used to exchange data. An extended LTE protocol stack and eNB functionality is stated in [10] and now listed:

- Locating and checking D2D peers positions in cells (for devices being in the Active mode)
- Identification of an active D2D communication.
- Authorization and allocation of dedicated D2D discovery resources for the D2D pairs
- Mode selection by the means of specific radio measurements on the cell (global congestion check on the cell) and with the UEs
- Dedicated D2D Radio bearer establishment with D2D peers and resource allocation for the D2D communication

IV. Proposed approaches and solutions to reduce Latency

V. Future Improvements for 5G

VI. Conclusion

REFERENCES

[1] IWPC white paper, Mobile Multi Gigabit (Mogig) Wireless Networks And Terminals âĂŞ 5000x Working Group, April 2,2014 http://iwpc.org/WhitePapers.aspx#5000x METIS requirements, presentations by Samsung, Intel, Ericsson, 5GNow, etc. etc.

[2] "Scenarios, requirements and KPIs for 5G mobile and wireless system," ICT-317669-METIS/D1.1, METIS deliverable D1.1, Apr. 2013. [Online]. Available: https://www.metis2020.com/documents/deliverables

[3] Latency in 5G, 4G, Juni, 9,2014 âĂŞ Don Brown & Stephen Wilkus http://5gnews.org/latency-5g-legacy-4g/

[4] D2D-based V2V Communications with Latency and Reliability Constraints - Globecom 2014 Workshop - Ultra-Low Latency and Ultra-High Reliability in Wireless Communications

[5] C. Lottermann, M. Botsov, P. Fertl, and R. Mullner, âĂIJPerformance evaluation of automotive off-board applications in LTE deployments,âĂİ in IEEE Vehicular Networking Conference (VNC), 2012.

[6] Wireless Device-to-Device Communications and Networks âĂŞ By Lingyang Song, Dusit Niyato, Zhu Han, Ekram Hossain - https://books.google.de/books?id=d2saCAAAQBAJ&lpg=PA132&ots=_IjrktfGW4&dq=hop%20gain%20lte&pg=PA162#v=onepage&q=interference&f=false

[7] Latency Improvements in 3G Long Term Evolution T. BlajiÄĞ, D. NoguliÄĞ, M. DruÅ¿ijaniÄĞ

[8] Radio Protocol Architecture http://www.tutorialspoint.com/lte/lte_radio_protocol_architecture.htm

[9] S. Sadr, A. Anpalagan, and K. Raahemifar, âĂIJA novel subcarrier allocation algorithm for multiuser OFDM system with fairness: UserâĂŹs perspective,âĂİ in IEEE Vehicular Technology Conference (VTC), 2007, pp. 1772âĂŞ1776.

[10] Hybrid model for LTE Network-Assisted D2D communications, Thouraya TOUKABRI GUNES, Steve TSANG KWONG U and Hossam AFIFI, Orange Labs, Issy-les-Moulineaux, France