Latency Improvements in 3G Long Term Evolution

T. Blajić, D. Nogulić, M. Družijanić

Mobile Solutions
Ericsson Nikola Tesla d.d.
Krapinska 45, PO Box 93 Zagreb, Croatia
Telefon: +385-1-365 3702 Fax: +385-1-365 3219 E-mail: tomislav.blajic@ericsson.com

Abstract - To ensure continued competitiveness of 3G technology for the next decade, 3GPP is establishing evolution plans that are following introduction of HSDPA and Enhanced Uplink. The new major step is known as 3G Long Term Evolution (LTE) and is based on several promising technologies, like OFDM and MIMO, involving also the System Architecture Evolution (SAE).

Major performance goals addressed by LTE are significantly increased peak data rates, reduced latency, spectrum efficiency, together with lower cost and complexity. This paper presents main mechanisms applied for improving latency in new, evolved system, both in control and user plane.

I. INTRODUCTION

Latency has a major influence on user experience. In particular, conversational services, such as VoIP and video telephony, require low latency. Other services that benefit from low delay are gaming and applications with extensive handshaking, such as e-mail. As user requirements for support of new services in mobile environment is constantly increasing, this is pushing continuous evolution of mobile technologies.

It is difficult to substantially improve latency without reducing the transmission time interval (TTI). The roundtrip time (RTT) in advanced GSM/EDGE networks with is around 150ms. Current WCDMA networks further on improve latency - with HSDPA it is around 80 ms, introduction of Enhanced Uplink (HSUPA) reduce it further on. Further evolution of 3G is aiming (among other targets) to improve latency to levels around 5 ms.

Main mechanisms for latency reduction include shorter Time Transmit Intervals (TTI), and faster feedback mechanisms (HARQ procedures).

Reducing the TTI improves latency substantially and immediately. TTI in EDGE is 20 ms, WCDMA brings TTI to 10 ms, while in HSDPA it is brought to level of 2 ms.

To help the transmitter better understand the radio environment, feedback information is sent via the radio link control (RLC) protocol over the air. The RLC protocol typically runs in acknowledged mode, which requires the retransmission of lost radio blocks. Although feedback information is crucial for efficient transmission over the radio interface, it is also time-consuming. The procedure requires the receiver to periodically send (on request):

- acknowledgements of radio transmissions;
- information about the current radio environment.

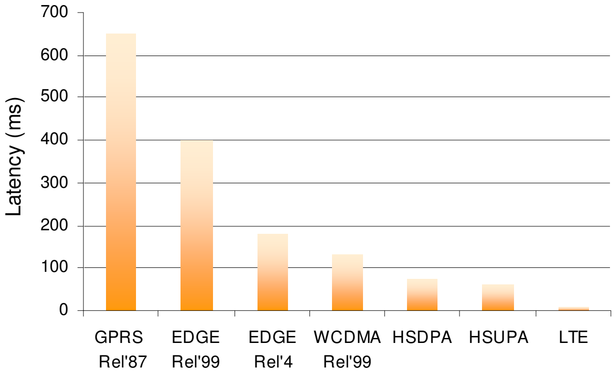

Figure 1. Latency performance in mobile networks

Faster feedback enables the transmitter to retransmit lost data earlier and makes radio transmission more efficient. By putting more stringent requirements on reaction times and by introducing support for immediate response to unsuccessful radio transmissions, one can ensure that lost radio blocks are retransmitted much earlier, which reduces latency. One can reduce latency even further by combining faster feedback with reduced TTI.

II. LATENCY ISSUES

As part of the IMS realtime services using HSDPA/HSUPA, VoIP and gaming over internet may become very important applications for the operators. To ensure the target performance and propose proper improvements of existing standards, it is necessary to investigate and define the performance requirements for this traffic. 3GPP defines [3] requirements specified in Table 1.

If VoIP over HSDPA is intended to become the primary mechanism for providing voice and/or video services, then it can be assumed that the requirements should not be degraded from the existing requirements shown above.

Requirements for realtime gaming are obviously very dependent on the specific game, but it is clear that demanding applications will require very short delays. Consistent with demanding interactive applications, a maximum of 75 ms end-to-end (E2E) one-way delay is proposed here and leads to a ping time of 150 ms.

Jitter has been shown to be problematic, but not in a truly quantifiable way and it is proposed that jitter should be minimized to as little as possible, potentially by algorithms in the application rather than in the network.

TABLE 1.
END-USER PERFORMANCE EXPECTATIONS - CONVERSATIONAL / REAL-TIME SERVICES

Medium	**Application**	**Degree of symmetry**	**Data rate**	**e2e one-way delay**	**Delay variation within a call**	**Information loss**
Audio	Conversational voice	Two-way	4-25 kb/s	< 150 msec preferred < 400 msec limit	< 1 msec	< 3% FER
Video	Videophone	Two-way	32-384 kb/s	< 150 msec preferred < 400 msec limit Lip-synch < 100 msec		< 1% FER
Data	Telemetry	Two-way	<28.8 kb/s	< 250 msec	N.A	Zero
Data	Realtime games	Two-way	< 60 kb/s	< 75 msec preferred	N.A	< 3% FER preferred < 5% FER limit
Data	Telnet	Two-way (asymmetric)	< 1 KB	< 250 msec	N.A	Zero

A. Gaming aspects

The following three major performance parameters have been identified to be sufficient for defining the gaming performance requirements:

- End-to-end delay
- Jitter
- Application packet loss

Given the wide variety of games available, they can be separated into the following four categories in which the corresponding tolerable gaming performance have been described with the use of the above proposed performance parameters:

- First Person Shooter (FPS), Racing – fast user response, many online players at once, highly dynamic
  - Up to 150ms end-to-end delay may be acceptable
  - 10ms jitter may be expected to be critical for FPS
  - Up to 5% packet loss is acceptable
- Real Time Strategy (RTS), Simulations – slightly lower response required, slower gameplay, handful of players in a single game
  - 250ms-500ms end-to-end delay may be acceptable
  - Requirements for jitter are undefined
  - 1% packet loss with 150ms delay may be acceptable
- Massive Multiplayer Online Role Playing Games (MMORPG) – Persistent games, highly variable scenarios, many hundreds of players online at once, many tens in a given situation
  - Several packets every ms, Latency << 350ms
  - Depending on the time and game contents, the required data rate ranges between 8kbps- 24kbps
  - 10% packet loss may be acceptable if latency is low;
- Non-real Time Games (NRTG) – e.g., chess, backgammon, cards etc.
  - Zero packet loss, which can be achieved by retransmission methods

It is understood that, within these four gaming categories, the requirements for delay, jitter and packet loss are different depending on the games and on the expectations of the player, but the typical range of the requirement attributes for real time gaming have been observed as following:

- Packet loss 0.1% - 5%
- Latency (e2e) 75ms – 250ms
- Data rate (5kbps- 60kbps)

It is clear from the values above that the 3GPP definition is considerably stricter than necessary in terms of packet loss in most cases. However, in terms of latency, the requirements on the RAB delay for best performance are stretched since E2E delays >75ms will lead to a perceived drop in performance from the user if the game requires the fastest response – some examples[2] depending on the game perspective and interaction model, through avatar (first or third-person view) or omnipresent, are shown on Figure 2.

Improved latency behavior in future 3G evolution will definitely be required for high-quality user experience in online gaming.

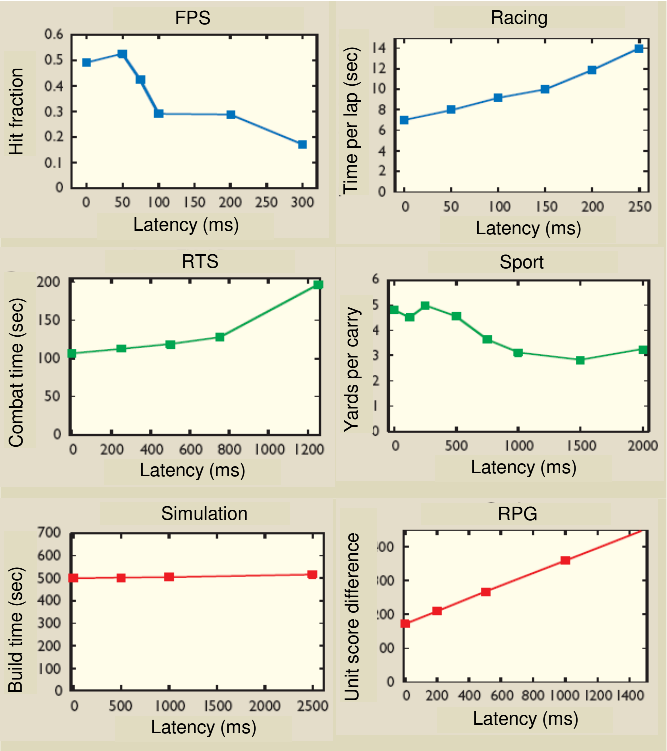

Figure 2. Player performance versus latency for individual actions.

B. Delays for PS and CS Connection Procedures

Another aspect of latency in mobile networks can be related to transition between different UE states during CS or PS connection, that affect user perceived service quality.

In Rel-99, UMTS introduced a dedicated channel (DCH) that can be used for CS and PS connections when UE is in CELL_DCH state. In addition to CELL_DCH state, other states were introduced - CELL_FACH state where signalling and data transmission is possible on common channels (RACH and FACH) and CELL_PCH and URA_PCH states, where the transmission of signalling or user data is not possible but enables UE power savings during inactivity periods maintaining the RRC connection between UE and UTRAN and signalling connection between UE and PS CN. The introduction of the CELL_PCH and URA_PCH states, the need of releasing the RRC connection and moving the UE to Idle mode for PS connections was removed and thus the Rel-99 UTRAN can provide long living Iu-connection PS services.

On the other hand, when UE is moved to CELL_PCH or URA_PCH state, the start of data transmission again after inactivity suffers inherent state transition delay before the data transmission can continue in CELL_DCH state. As new packet-oriented features like HSDPA and HSUPA in Rel-5 and Rel-6 UMTS systems respectively provide higher data rates for both downlink and uplink in CELL_DCH state, the state transition delay has been considered to be significant and negatively influencing the end user experience.

In addition to RRC state transition delay, the radio bearer setup delay to activate new PS and CS services has been seen as problematic in UMTS, due to signalling delays on CELL_FACH state where only low data rates are available via RACH and FACH, and due to activation time used to synchronize the reconfiguration of the physical and transport channel in CELL_DCH state.

To secure future competitiveness of UMTS and enhance the end user experience even further, the delay optimization for procedures applicable to PS and CS connections work is targeted to reduce both setup times of new PS and CS services and state transition delays to, but still enable, excellent UE power saving provided by CELL_PCH and URA_PCH states.

The agreed modifications can be summarized as: introduction of enhanced support of default configurations, reduced effects of the activation time, and utilization of HSPA for signalling. From Rel-6 the signalling radio bearers (SRBs) can be mapped on HSDPA and HSUPA immediately in RRC connection setup and default configurations can be used in radio bearer setup message and RRC connection setup message in a more flexible way.

The utilization of default configuration and mapping of the SRBs on HSDPA and HSUPA will reduce message sizes, activation times, and introduce faster transmission channels for the signalling procedures, thereby providing significant enhancement to setup times of PS and CS services compared to Rel-99 performance.

In the 3GPP Rel-7 time frame, the work will study methods of improving the performance even further, especially in the area of state transition delays. As the work for Rel-7 is less limited in scope of possible solutions, significant improvements to both RRC state transition delays and service setups times are expected.

III. 3G LTE

To meet challenge of further increase of user demands and competition from new radio access technologies in longer time perspective, 3GPP has initiated the study item Evolved UTRA and UTRAN [4]. Aim of this study for the long-term evolution of 3G (3G LTE) is to achieve major leaps in performance in order to improve service provisioning and cost reduction. These targets [5] include:

- Peak data rates up to 100 Mbps for the downlink direction and 50 Mbps for the uplink direction.
- Mean user throughput improved by factors 2 and 3 for uplink and downlink respectively.
- Cell-edge user throughput improved by a factor 2 for uplink and downlink
- Improved coverage (high data rates with wide-area coverage) and capacity (threefold capacity compared to current standards)
- Uplink and downlink spectrum efficiency improved by factors 2 and 3 respectively.
- Significantly reduced latency in the user plane in the interest of improving the performance of higher layer protocols (TCP) as well as reducing the delay associated with control plane procedures (session setup)
- Reduced cost for operator and end user
- Spectrum flexibility, enabling deployment in many different spectrum allocations.

A. Key technologies for Long Term-Evolution

The ability to provide high bit rates is a one of key measures for LTE. Multiple parallel data stream transmission to a single terminal, using Multiple-Input-Multiple-Output (MIMO) techniques, is one important component to reach this. Beamforming antennas can also be used to improve coverage of a particular data rate and to increase the system spectral efficiency.

Regarding choice of radio access technique, by applying Adaptive Multi-Layer Orthogonal Frequency Division Multiplexing (AML-OFDM) in downlink, system will be able to operate at different bandwidths in general providing large bandwidths for high data rates. Varying spectrum allocations, ranging from 1.25 MHz to 20 MHz, are supported by allocating corresponding numbers of AML-OFDM subcarriers. Operation in both paired and unpaired spectrum is possible as both time-division and frequency-division duplex are supported by AML-OFDM. OFDM is also suitable for broadcast services, as OFDM receiver in terminal exploits broadcasting information from several base stations as mulipath propagation.

In uplink, a key requirement is set on power-efficient user-terminal transmission to maximize coverage. To reach this, Single-Carrier Frequency-Division Multiple Access (SC-FDMA) with dynamic bandwidth is a good choice. In order to achieve intra-cell orthogonality, the base station assigns a unique time-frequency interval to the terminal for the transmission of user data. This is done for each time interval. The users are separated primarily by time-domain scheduling; however, if the terminal has a limited transmission power or not enough data to transmit, also frequency-domain scheduling is used.

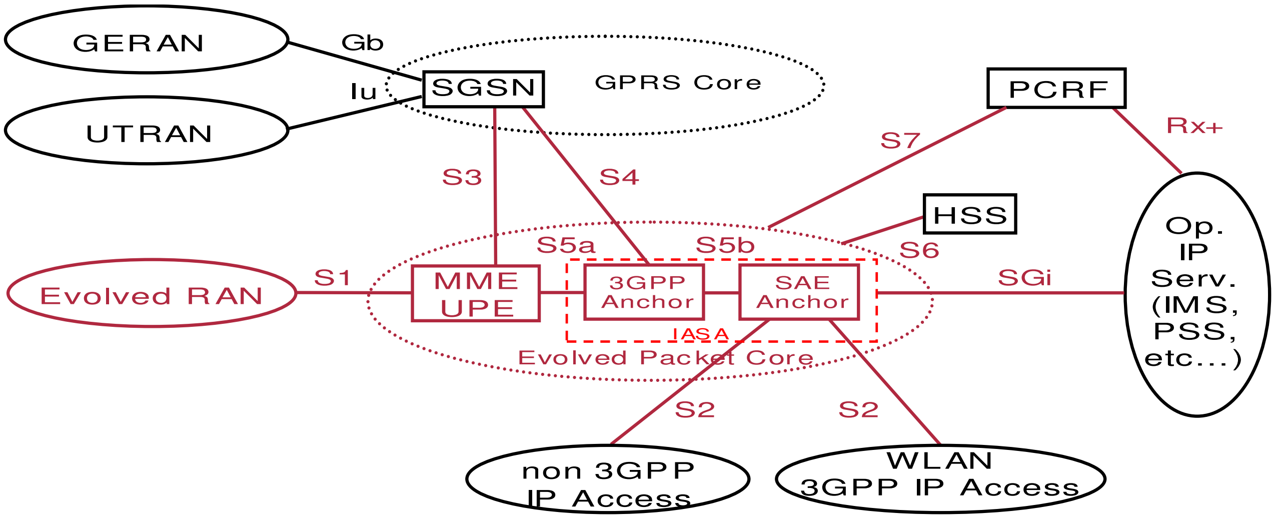

Figure 3. Logical high level architecture for the System Architecture Evolution (SAE)

Other important building blocks for the long-term 3G evolution are system architecture optimized for packet services, and evolved quality of service, QoS, and link-layer concepts.

B. System Architecture Evolution

Given the requirements to reduce latency and cost, it makes sense to consider a system architecture that contains fewer network nodes, because this reduces the overall amount of protocol-related processing, number of interfaces, and cost of interoperability testing. Fewer nodes can also result in easier radio interface protocols optimization, e.g. by merging some control plane protocols. Shorter signalling sequences result also in more rapid session setup.

LTE concepts of simplified system architecture are closely related to the 3GPP work on System Architecture Evolution (SAE). SAE currently defines [6] base line high level architecture for the evolved system (Figure 3), defining new elements of Evolved Packet Core (ECP) network.

Two new entities are introduced, which are in LTE terminology commonly referred as Access Gateway (aGW):

- Mobility Management Entity (MME): manages and stores UE context (for idle state: UE/user identities, UE mobility state, user security parameters). It generates temporary identities and allocates them to UEs. It checks the authorization whether the UE may camp on the TA or on the PLMN. It also authenticates the user.
- User Plane Entity (UPE): terminates for idle state UEs the downlink data path and triggers/initiates paging when downlink data arrive for the UE. It manages and stores UE contexts, e.g. parameters of the IP bearer service or network internal routing information. It performs replication of the user traffic in case of interception.

Two functional entities that anchor the user plane are also defined (still it is not decided if they will be co-located with the MME/UPE, or separate nodes). 3GPP Anchor manages mobility between the 2G/3G access system and the LTE access system, while SAE Anchor affects mobility between 3GPP access systems and non-3GPP access systems.

C. Latency concepts in 3G LTE

As it can be seen from LTE targets, aim is set to address latency from 2 different aspects - control plane and user plane latency. 3GPP has defined this in TR 25.913 [5].

Control plane latency is related to state transition times, where significant reduction is expected for (Figure 4):

- transition time from a camped-state (such as Release 6 Idle Mode) to active state (such as Release 6 CELL_DCH) where user plane is established - excluding downlink paging delay and NAS signalling delay this should be less than 100 ms (it is also assumed that UE is already attached and hence no Authentication and Attach procedures are required)
- transition time between inactive (dormant) state (such as Release 6 CELL_PCH) and active state (such as Release 6 CELL_DCH) - aim is to achieve less than 50 ms (excluding DRX interval)

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.

Specifications aim on user-plane latency of less than 5 ms in unloaded condition (single user with single data stream) for small IP packet (e.g. 0 byte payload + IP headers). It can also be expected that LTE bandwidth mode may impact the experienced latency

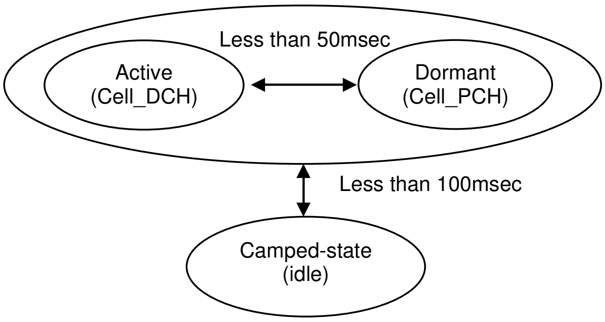

Figure 4. Examples of state transition

As 3GPP system architecture (SAE) is still not settled, including the RAN and core network functional split, network entities between which the user-plane latency requirement applies, will finally be defined at a later stage of standardization.

The ability to reach high bit rates is highly dependent on short delays in the system and a prerequisite for this is short sub-frame duration. Consequently, the LTE sub-frame duration is set as short as 0.5 ms in order to minimize the radio-interface latency

D. Latency in control-plane

With the current framework definition for LTE/SAE, the radio control plane protocol stack previously resident in the RNC is now located in the eNode B; the user plane protocol stack is split between the AGW and the eNode B. Most of the Rel-6 CN protocol functionality is expected to reside in the AGW/CPE.

The following elements contribute to the control-plane latency:

- Transmission delay
- Retransmissions for reliable transfer
- eNodeB/UE L1/L2/L3

The overall latency for change in state from LTE_IDLE to LTE_ACTIVE will be affected by reduction of the number of messages exchanged between the UE and the NW before data transfer can be initiated. Possible two optimizations [7] for LTE include proposal that piggy-back transfer of NAS messages would reduce overall latency significantly and the concept of a default Radio bearer. Considering this, one potential resultant message flow sequence could be the following (UE initiating data transfer):

1. The UE transmits the RRC connection Request.

2. The connection request results in the eNode B requesting transfer of the UE related context transfer from the AGW.

3. The AGW responds with the initiation of the security procedures and transfer of context.

4. The UE sends back the L3 ACK along with the security complete message to be forwarded to the AGW. There will be a finite delay coming from scheduler action before the UE gets scheduled and is able to transmit/receive data

Following assumptions can be used to estimate each leg of message flow that contributes to the latency (Table 2):

- It is assumed that each AS and NAS message or the piggy back combinations can be transmitted in one sub-frame. All processing times are therefore in multiples of 0.5 ms.
- Layer 1 processing times of 2 x 0.5 ms is assumed for both transmission and reception. Additional L2/L3 processing of 1 sub-frame is considered at each of the nodes and UE.
- N=6 SAW HARQ is assumed on the radio side, with a HARQ retransmission probability of 30%.
- A zero delay over the S1 control-plane is assumed.

The random access procedure can be a significant contributor to the access delay. For the LTE_IDLE to LTE_ACTIVE state, there is only one random access procedure that needs to be initiated – the transfer of the RRC Connection request message at the very outset. A delay of 10 ms is assumed. The actual delay will be dictated by the persistence, RACH channel design and other parameters

TABLE 2.
CONTROL-PLANE LATENCY ESTIMATION
FOR LTE_IDLE TO LTE_ACTIVE STATE TRANSITION

UE – eNode B		**eNode B – aGW**	
TX L3/L2 processing	1 ms	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	**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**		

With the above assumptions, the time interval for message flow shown on Figure 5. is in order of 31 ms. After this procedure Node B can schedule the UE for DL and UL transmissions. As this may result in another 2-3 ms delay, an overall delay of 34 ms is achievable.

Currently, a dormant state has not been defined for the MAC protocol layer. RRC has only two states, IDLE and CONNECTED. In the RRC CONNECTED state, MAC is in ACTIVE state. The UE may enter DRX in which case it would wake up at pre-configured times. If the UE has lost uplink synchronization while in DRX, the UE will need to transmit a message using the non-synchronized RACH. A delay of around 10 ms is estimated before it receives a TIMING ADVANCE message along with resources for uplink transmission. Following this, the UE sends scheduling information leading to the network providing it a resource allocation 2-3 ms later. No HARQ mechanism is assumed for the random access procedure. This provides a total delay of around 12-13 ms

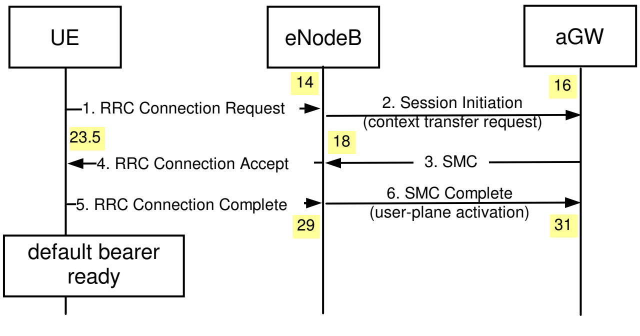

Figure 5. Delay Snapshot for idle to active LTE transition

E. Evaluation of user-plane latency

According to 3GPP user-plane latency definition, it could be argued what the RAN edge node refers to - eNodeB or aGW. Evaluation presented here [8] indicate that there won't be a large difference in approach, making it feasible to reach the user plane latency requirement of 5 ms.

For the evaluation, it is assumed that the network is operated with low load. The "unload condition" assumption of a single user with single data stream implies no scheduling delays due to other users and data streams. From latency point of view this justifies the assumptions of immediate scheduling both at radio interface and eNB-aGW-interface. This implies that queuing delays for both can be excluded. Furthermore, as the radio resource usage is not an issue in a single user case, it can be assumed for the baseline evaluation that sufficient block error probability can be achieved without RLC and possibly even without HARQ retransmissions. Finally, the "small IP packet" assumption implies that the packet fits in one subframe and no segmentation is used

The results are therefore a lower bound. Delay values in loaded scenarios might be significantly higher, in particular for low priority traffic, because queuing and scheduling delays become dominating in such scenarios. Further it is assumed the UE is in LTE_ACTIVE state and is synchronized to the network, i.e. no extra delay due to random access procedure.

Based on the assumptions above, following relevant delay components for user-plane latency evaluation can be identified:

- 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 latency values for the radio interface are dominated by the TTI and the UE-eNB latency is estimated to 2 ms.

For the S1 interface it is assumed that the associated latencies are dominated by the propagation delay. According to the propagation speed in copper cables of 200,000 km/s, a distance between two nodes of 200 km results in a latency of 1 ms. Obviously, the network topology choice has therefore a significant impact on the system latency.

Finally, 0.5 ms processing latency is assumed for the user plane processing in the aGW.

In total this give a latency from the UE to the aGW of 3.5 ms. The estimated latency components for the user plane in LTE are summarized in Figure 6.

Various values of HARQ retransmission and eNB-aGW delays might influence delay values. With sufficiently low HARQ retransmission (< 30%) and eNB-aGW (< 1ms) delays, the specifications enable targeted LTE user-plane latency of less than 5 ms in unload condition for small IP packet.

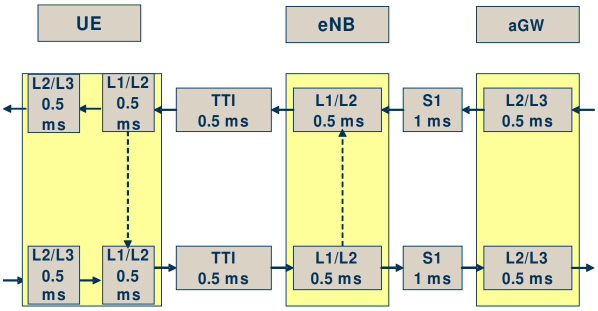

Figure 6. User plane latency components in LTE

IV. CONCLUSION

3GPP is in the process of defining the long-term evolution (LTE) for 3G radio access, in order to maintain the future competitiveness of 3G technology. The LTE physical layer has benefits in terms of broadcast, spectrum flexibility, and the possibility for frequency domain adaptation. The LTE concept is not limited to the physical layer - architecture and higher layer protocol enhancements are also included through System Architecture Evolution (SAE).

Along with other main targets of this evolution (increased data rates, improved spectrum efficiency, improved coverage), latency aspects are becoming more and more important for support of new interactive services like VoIP or online gaming. Evaluated control-plane delays for IDLE to ACTIVE and dormant to ACTIVE transitions in LTE are well within the targets of 100 and 50 ms. Analysis of the user plane latency in LTE also shows that in a low loaded network it is feasible to reach the user plane latency requirement of 5 ms.

REFERENCES

[1] E. Dahlaman at al., "The 3G Long-Term Evolution – Radio Interface Concepts and Performance Evaluation", *IEEE VTC,* spring 2006.

[2] M. Claypool, K. Claypool., "Latency and Player Actions in Online Games", *Communications of the ACM,* Nov. 2006.

[3] 3GPP TS 22.105 V8.2.0, "Services and service capabilities", Dec. 2006.

[4] 3GPP TR 25.814 V7.0.0, "Physical layer aspects for evolved Universal Terrestrial Radio Access (UTRA)", May 2006.

[5] 3GPP TR 25.913 V7.3.0, "Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN)", March 2006.

[6] 3GPP TR 23.882 V1.3.0, "3GPP System Architecture Evolution: Report on Technical Options and Conclusions", July 2006.

[7] Motorola, "C-Plane Latency – Analysis", *3GPP TSG-RAN WG2 Meeting #53*, May 2006.

[8] Ericsson, "Concept evaluation of user plane latency in LTE", *3GPP TSG-RAN WG2 Meeting #53*, May 2006.