Low-Latency in Wireless Communication

Navid NIKAEIN[1], Raymond KNOPP[1,] Antonio Maria CIPRIANO[2]
Srdjan KRCO[3], Igor TOMIC[3], Philipp SVOBODA[4], Markus LANER[4], Eric LARSSON[5], Yi WU[5], Manuel GARCIA FUERTES[6], Janie BAÑOS[6], Nenad ZELJKOVIC[7], Djordje MAROVIC[7]

[1]*Eurecom, 06504 Sophia Antipolis, France*
[2]*Thales Communications, Colombes, 92704, France*
[3]*Ericsson d.o.o, Novi Beograd, 11070, Serbia*
[4]*Technical University of Vienna, Vienna, Austria*
[4]*Linköping University, Linköping, Sweden*
[6]*AT4WIRELES, Malaga, 29590 Spain*
[7]*Telekom Srbija a.d., Beograd, 11070 Serbia*

Abstract: This paper focus is on access-layer technologies targeting low-latency robust and spectrally-efficient transmission in a set of emerging application scenarios. Two basic types of wireless networks are considered, namely long-term LTE-Advanced cellular networks and medium-range rapidly deployable mesh networks. In cellular networks, research is focused on transmission technologies in support of gaming services which will undoubtedly prove to be a strategic revenue area for operators in the years to come. We also consider machine-to-machine (M2M) applications in mobile environments using sensors connected to public infrastructure (in train, buses, stations, etc.).

Keywords: Latency, LTE, LTE-A, M2M Communication Scenario, Interactive Gaming, Real-Time Application, System Architecture

1. Introduction

Latency is a major factor influencing the experience of user, machine, and any form of communicating from the application used. The majority of applications will not benefit much from lower latency than that offered by LTE, or at least users would likely not be willing to accept increased subscription rates for this feature. Important exceptions to this, however, are interactive multi-player gaming applications which, from an operator's perspective, represent a very strategic application area with respect to revenue potential. In addition, for realtime machine-to-machine (M2M) applications in the area of intelligent transport system, remote monitoring and health will be also requiring very low latency and are become the main focus of many mobile and IT operators and vendors as a new revenue opportunity.

In the following subsections, first, we provide the basic notion of latency followed by a set of low latency application scenarios for both M2M and gaming. Then, the M2M and gaming traffic characteristics are analyzed. Finally, a set of measurement is carried out to study the delay and traffic for online gaming scenario, and some techniques are presented to reduce the latency at the access layer.

2. Notions of Latency

One of the important design objectives of LTE/LTE-A (and to some extent HSPA) has been to reduce the network latency. Network latency consists of both c-plane and u-plane latency.

In the 3GPP Literature [4], the c-plane latency can be defined as the time taken by the first packet to successfully reach the receiver reference point. In LTE/LTE-A, the c-plane latency is defined as a transition time between two states, IDLE or DRX to ACTIVE. Typically, in LTE/LTE-A the transition time from IDLE to ACTIVE state should be less than 100ms, and from DRX to ACTIVE state depends on the DRX cycle. The user plane latency, also known as transport delay, is defined as the one-way transit time between a packet being available at the IP layer of the sender and the availability of this packet at the IP layer of the receiver. In LTE/LTE-A, this latency is defined between the UE and EPC edge nodes. LTE/LTE-A specifications target the user-plane latency of less than 5ms in unloaded condition (single user with single data stream) for a small IP packet with no payload.

However, latency can also be interpreted in terms of the efficiency of very low-layer procedures allowing for time/frequency synchronization, identification/authentication, channel setup time, channel interleaving, channel code block length, etc. These are all fundamentally related to signaling overhead and channel-code complexity (i.e. block-length) in the access stratum. The majority of wireless systems, including LTE, are designed for a continuous flow of information, at least in terms of the time-scales needed to send several IP packets (often large for user-plane data) containing information and such overhead is manageable. In some evolving M2M application scenarios (e.g. remote "smart" metering) packets are short and small in number and extremely low duty-cycle, which from a system throughput perspective represents a vanishing data rate. In such low spectral-efficiency applications (seen by the application not the aggregate spectral efficiency), the signaling overhead latency translates directly into energy-loss, due to the fact that the whole embedded system is the sensing device is powered-up during the synchronization/training procedure prior to sending/receiving a short packet. While for a particular M2M or sensing node this energy-loss can be negligible, the aggregate cost due to the unbounded number of nodes could prove to be significant from a network standpoint. This clearly calls for a more detailed definition of latency in the presence of M2M traffic with conventional user traffic, and coupled with the potential of a rapid increase in the number of machines connected to cellular infrastructure in the coming decades.

3. Low Latency Application Scenarios

In this section, we focus on two types of applications: real-time machine to machine communication and interactive gaming [7].

3.1 – Realtime Machine-to-Machine Communication

Although a large variety of M2M application scenarios with heterogeneous requirements and features exists, they can be classified into two main M2M communication scenarios as defined in [5], communication of M2M devices with M2M servers/users and communication between the M2M devices.

At the present time, the most interesting applications from the commercial point of view are related to smart electricity, automatic water and gas meters reading. However, the M2M application space is vast and includes security, health monitoring, remote management and control, tracking and tracing, intelligent transport systems,

distributed/mobile computing and gaming, industrial wireless automation, and ambient assisted living.

3.1.1 AutoPilot

This scenario includes both vehicle collision detection and avoidance (especially on highways) and how urgency actions are taken in case of an accident. It is based on a M2M device equipped with sensors embedded in cars and surrounding environment and used in automatic driving systems. These M2M devices (cars, road sign units, highway cameras) send information to a backend collision avoidance system used. The backend system distributes notifications to all vehicles in the vicinity of the location of the collision, together with information required for potential actuation of relevant controls in the affected cars. In all the receiving cars the automatic driving systems based on the received information take over the control fully or partially (brakes activated, driving direction changed, seating belts tightened, passengers alerted etc). If there is no such system in a car, the driver is notified and instructed. Also, depending on the proximity of the accident, cars receive different commands, i.e. the cars which are closer to the place of the possible collision are getting immediate commands for the actuators, while the cars which are further away from this place get driver notifications only.

3.1.2 Virtual Race

One example of the many possible M2M games is the virtual race (e.g. virtual bicycle race using real bicycles). The opponents are on different locations, possibly many kilometers away. At the beginning, the corresponding length of a race is agreed (i.e. 10 km or 20 min) between the peers. The measurements are taken by sensors (GPS, temperature, humidity, speed, terrain configuration etc.) and are exchanged between the opponents. They are used by the application to calculate the equivalent positions of the participants and to show them the corresponding state of the race (e.g. “you are leading by 10 m”). The number of competitors may be more than two, and all competitors must mutually exchange information, and the applications must present all participants the state of other competitors. For a large number of competitors (hundreds or more), a corresponding application server must be used. During the race they are informed about the place and the distances from each other (e.g. “you are the 3rd behind the 2nd by 10 m and leading before the 4th by 15 m”).

3.1.3 Sensor-based Alarm or Event Detection

Many categories of applications exist or will be reasonably implemented in the future. In some applications, sensors infrequently deliver a small amount of data: e.g. high risk transportation, meteorological alerts, stability of buildings, critical parameters in plants, etc. Of course the type of power supply (if the sensor is always on or not), density and other parameters depend on the application. Another type of application is event detection requiring fast reaction. An example is the detection of pressure drop through the pipelines (gas/oil); this critical information should be sent immediately to the control centre in order to prevent potential accidents. In the field of surveillance and security, discrete sensors which should stay undetected can enable interesting applications too. Examples of this type of applications can be intrusion detection sensors, or automated network of surveillance camera (with or without motion or pattern detection, mounted or not on robots, for instance), which send periodic reports to and interact with the control center, possibly in a completely automated way, until a critical event requiring the human intervention is detected. Depending on the type of applications, certain cases may require the deployment

of proprietary networks, or they may be run on top of a standard LTE/LTE-A network or of a mesh network deployed for a specific need. Only the operational context may decide of the exact network architecture.

3.1.4 Team Tracking

Team Tracking (TT) applications aim at monitoring the position of several nodes in a given environment for situation awareness and consequent action scheduling. One typical application of the TT scenario is the monitoring of firemen or policemen in a given area (e.g. building, stadium) during an operation. In this context, it is crucial to have an up-to-date picture of the situation, to allow reactivity and manage the team considering the positions and the associated risks. A constant monitoring is usually obtained thanks to a positioning system and the connectivity among the nodes and the control center. This kind of application is typically run on ad hoc mesh networks deployed for specific needs, hence the control center may physically collocated in one of the node of the network or it may be physically far away. In the latter case, one node of the network offers connectivity to an IP backbone in order to reach the control center.

3.2 –Gaming

At present time the evolution of processing power at the end terminal has lead to a point where even complex online games can be played anywhere even by highly mobile users, e.g., in public transport units. Even if this step can be considered as normal evolution, it is important to note that it introduces a new class of real-time applications into the world of mobile cellular wireless networks.

The core aspect of any gaming scenario is the fast inter-active nature of the application: users interact with other users. The nature of this interaction can be of different kind, like cooperative or challenging. The gaming applications span a wide range from racing simulations, over real-time strategy to first person-shooters. In general the classic applications challenge the reaction time of the users and therefore rely on low-delay access technologies. The logical network architecture is not necessarily limited to any specific setup like fully meshed nodes or server-client setups.

In contrast to most other applications, gaming traffic patterns strongly correlate with user interaction at the terminal side. A traffic stream from such a source is a function of participating users, their properties and so on. This is more than some data packets exchanged between two nodes.

3.2.1 Online Gaming

Online gaming is an application that evolved from multiplayer games. In multiplayer games two or more players enter the same game either in competition or in cooperation mode. The early realizations were realized at the same machine; later the players took part on separate computers connected via a network. In general the communication between the nodes in the game can be established over different networks and in different manners. In [6] the authors show that the games most sensitive to delay are the First Person Shooters (FPS) and sport games, e.g., car racing.

The “Online Gaming” setup is a classical multiplayer game in which two or more persons compete for a given time. Each game has a defined start and end. At the end all the scores are calculated and presented to all the participants. A high delay will impact the gaming experience of the user on one hand and the resulting score on the other hand [6]. These applications will typically take place in cellular mobile networks.

New on-line games will move latency requirements dramatically down. Low latency is especially critical for an avatar model of Online Games with high precision weapons (massively multi-user on-line first-person shooter).

3.2.2 Gaming on Sport Events

At present online gaming application take purely place in fictional and created worlds. However with increasing number of high performance handsets one could think of online games as a part of the real world. This is targeted by this scenario we call "Gaming on Sport Events" in this paper. Now imagine instead of competing on a virtual score the participants are in the audience of a big and exciting sport event. The goal of the game is the prediction of the next major event, like:

- Tennis: who will win the next point, will the next serve be an ace?
- Basketball: will next free-throw attempt be successful?
- Soccer: will the goal be scored from a free-kick?

It is very clear that low latency is critical due to real time context. In such scenario a high number of users in the same location (few radio cells) are expected. Furthermore, the same game can be played by TV audience. Then, the number of players can be much higher. In that case the application server has to process many more users. In any case, the traffic load can lead to increased delays. Such a setup can run either on top of LTE or LTE/A networks or as a self organizing meshed network in case of limiting the participants to the audience only.

4. Traffic Analysis

In this section, we analyze the traffic characteristics for machine-to-machine communication as well as for gaming application [7].

4.1 – Machine-to-Machine Communication

Today, mobile networks are dimensioned using standard mobile wireless network traffic models, which are based on a typical subscriber behaviour expressed in typical time spent using speech service, number of sent/received messages (SMS, MMS) and the amount of data subscriber is downloading. These traffic models do not take into account traffic generated by M2M devices, hence a new traffic model is required.

When modeling M2M traffic, two different patterns are expected. Most of the sensors will be in sleep mode, sending small amounts of traffic from time to time in order to notify the rest of the network that they are still alive. Some other M2M devices will generate traffic in bursts, so this type of traffic can be modeled with ON/OFF traffic model. Important parameters of traffic model that have to be considered are:

- potential number of devices and applications, density of devices
- periods of activity/sleep
- amount of useful data that is sent/received, for different type of applications
- overhead due to different types of signaling
- frequency and size of packets

The main expected differences in M2M traffic characteristics, comparing regular mobile network traffic patterns, are:

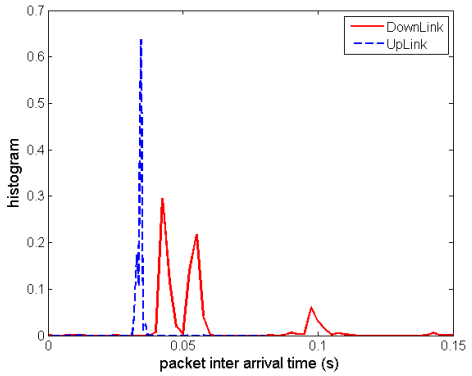
- M2M traffic will significantly increase number of parallel ongoing connections ("50 billion connected devices" by 2010 expected by Ericsson)
- M2M traffic will be more uplink consuming then downlink consuming, as it is expected that M2M devices in general have more information to send than to receive (confirmed in traffic model described in [3])

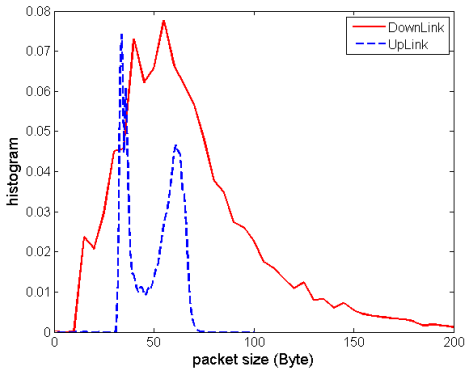
From a traffic analysis point of view it is interesting to note that some M2M device communicate using the TCP protocol. Therefore measurements will include some kind of interaction between the transport protocol and the actual access network under test. Any derived model has therefore be tested and analyzed for a possible presence of such an interdependency.

4.2 – Gaming

The online gaming applications that are most prone to end to end delays are FPS and racing games [6]. Therefore we selected well-known representative for these types of online games, namely: Team Fortress 2, DiRT2. In addition to this we selected also the FPS game open arena.

When analyzing online gaming traffic patterns, we expect the use of use the User Datagram Protocol (UDP) as transport protocol on top of IP. This protocol features no advanced flow control like e.g. the Transport Control Protocol (TCP) and therefore we can directly analyze the payload values of UDP which are the packet size and inter arrival time. In a standard client/server setup the traffic patterns for different online games are similar. This comes from the fact that data exchanged between server and client follows a given pattern. The downlink direction carries all the information about other players participating in the game. It is broadcasted by the server in a regular pattern to update the different users about the current status of the game. The traffic in downlink direction is a function of the number of real players participating in the game. A higher number of participants leads to more information that has to be transferred from the server to the clients. The packet size in downlink direction is therefore relative variable. In the uplink direction the client reports the actions of the local user to the server. As the single user can only execute a very limited number of actions in a timeslot, the uplink traffic consists of packets with a small spread in payload size, e.g., ranging from 40 to 70 Byte. The following two figures depict the packet size and the packet inter-arrival time for the open source game OpenArena. The game is based on the Quake 3 engine. The results follow the values expected for a FPS game as in [6].





The modeling of online gaming traffic can be done straight forward using packet size and inter arrival time as the underlying protocol is UDP. Therefore there is no interaction at the protocol layer. However, there is a correlation between the two values.

5. Delay and Traffic Measurement for Online Gaming

Delay measurements of Online Gaming were conducted in the existing network of Mobile Telekom Serbia (MTS) in order to get reference and idea of delay and traffic patterns in real network. The measurement campaigns recorded the different delay patterns found in the

access networks. As shown in the figure below, the technologies tested in this measurement are:

- ADSL Asymmetric Digital Subscriber Line,
- FTTB Fiber To The Building,
- WLAN Wireless Local Area Network,
- UMTS / 3G Universal Mobile Telecommunications System.

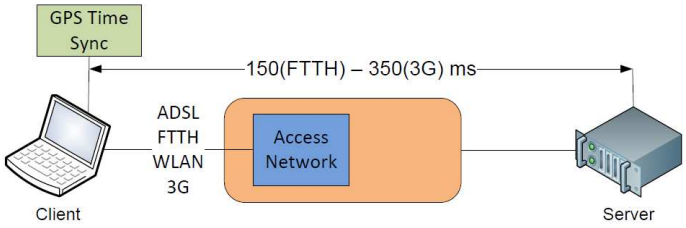


Figure 1: Measurement Setup for ADSL, FTTB, WLAN and 3G

The values obtained for the different technologies show a delay ranging from 150ms for the fast access types like FTTB to 350ms for 3G. The ping command via ICMP is used to measure the E2E delay. The server under test is the game server for the applications HalfLife and Colin Mc Rea Rally (two type of Online Gaming). The packet size is around 80Byte and nearly constant, while the inter packet time has a peak around 45ms between consecutive packets and a subsequent peak at 60ms. In contrast to this the uplink packet size histogram has three discrete spikes at 60, 75 and 90Byte, with an update rate of 30ms.

6. Latency Improvement in Access Layer

In this section, we present some possible access layer adaptation to lower the latency of the system for the considered applications. They include adaptive modulation and coding (AMC), scheduling, and hybrid analog/digital transmission [7].

6.1 AMC design

The current implementation of LTE has achieved a good flexibility and low latency in the management of HARQ and AMC when dealing with standard applications like VoIP or multimedia communications in a context of an unloaded network. However, the raise of M2M applications will probably highlight once again the old issue of resource management in presence of a high density of users (humans and, now, machines) possibly in conditions of high loaded. Moreover, the development of M2M application will probably set new challenges to HARQ and AMC, linked to the sparcity of the traffic, the presence of relay or multiple hops among devices, etc.

6.1.1 Policies for Sparse Traffic

Sparse traffic in some M2M applications raise challenges not only on the access and scheduling, but also on the choice of the best way to carry out information. Can spatial correlation be taken into account? What is the impact of having limited or null channel information? The influence of long sleep times may in fact be important in the choice of the right AMC and HARQ.

Certain applications (for example alarms) may need to send a limited number of information bits but with a very high reliability and in a strictly limited number of transmissions, due to strong latency constraints or battery life optimization. In such cases,

the most correct metric for measuring the delay introduced by HARQ could be the maximum peak delay values and not the average ones.

6.1.2 Policies for multihop network

The presence of multiple hops inside a cell of a future LTE-A network (or evolution) or on a rapidly deployable mesh network may have a strong impact on the definition of the most correct AMC and HARQ algorithms and procedures. For example, cooperative techniques may be considered: the expected gain is the increased robustness of the communication.
However, signalling and exact HARQ strategies need to be designed with the aim of minimizing latency. The impact on improved reliability on the throughput must also be controlled, and can make sense for certain M2M applications needing low user bit rates.
Finally, issues related to the time-frequency synchronization of the nodes and channel estimation must be considered when using cooperative techniques.

6.1.3 Efficient Feedback Signaling

Efficient feedback signalling for every king of PHY layer procedure (measurements procedures, ACK/NACK feedbacks, ranging and bandwidth request) will have a beneficial impact on the latency of the network. An improved design of this kind of messages can help in achieving the goal of the PHY layer procedures in harsh conditions (many devices wanting to access, limited number of opportunity in time and frequency due to battery constraint or to load, etc.). As a first example, in certain applications, sending with the ACK/NACK feedback also the amount of additional mutual information needed by the receiver can help the transmitter to limit the number of retransmissions.

6.2 – Scheduling

The study of MAC-layer scheduling policies and more generally channel-access is clearly of interest for M2M and certain gaming applications. In both application areas scheduling should be studied both from the point-of-view of the downlink and uplink. In the downlink the major issue is accommodating low-throughput and, more often than not, sporadic latency constrained traffic at the same time as conventional traffic.

On the uplink, the primary concern is channel access for sporadic traffic with minimal uplink signaling overhead. The latter is for energy concerns, in the sense that a low-throughput service should not require a significant energy overhead in order to maintain low-latency. In the extreme case, which could arise in remote sensing applications, one could clearly imagine the need for sending short packets with extremely low duty-cycle. Here a low-latency channel access protocol will keep energy consumption of the terminal device to a minimum, which in dense sensor networks is clearly a concern.

The trend in LTE which is culminating in the Release-10 standardization process is to reduce latency in the uplink channel access protocol. In the access stratum, this can occur with two-levels of granularity, basically depending on the activation state of the terminal. In the idle state (RRC_IDLE), the terminal must re-authenticate itself using the random-access procedure (to obtain a C-RNTI) which incurs a minimum-latency on the order of 16ms before being granted a transmission opportunity to upload a single packet. This value is significantly higher if the PRACH is configured with a high periodicity. This latency is difficult to circumvent and furthermore, the terminal must activate its transmission and reception circuitry during the access period, which is clearly not energy-efficient. It is clearly also a very inefficient access protocol for networks with a large number of low-throughput sporadic access nodes (e.g. sensor networks).

In the connected state (RRC_CONNECTED) the UE provides a scheduling request (SR) on the uplink control channel (PUCCH) resource in order to transmit a packet. Both in Release-8 and Release-10, the latency incurred by the channel access amounts to 6ms after the transmission of a SR. This is due to the reaction time of the eNB scheduler and the UL grant procedure via the downlink control channel (PDCCH). Aside from this constant latency, the remainder depends on the periodicity of the SR. In Release-8 LTE this can be at least 5 ms, but also much more and in Release-10 it is planned to reduce this to the strict minimum, namely 1ms. For SR periods larger than 5ms, the random-access procedure is more efficient for sporadic packet arrivals than through the SR and thus may prove to be a more suitable access technique for some M2M and gaming traffic flows when it comes to the UL. Moreover, a very low period SR is not suitable for very sporadic traffic sources, especially in dense M2M networks, since the signaling overhead quickly becomes comparable to the data throughput, which is clearly unacceptable.

As a result of the limitations in both of these UE states, it is worthwhile to consider new channel access techniques to accommodate sporadic traffic sources in LTE, especially if many such new devices start to use the networks. One avenue to address this would be to introduce new contention-based random-access.

6.3 - Hybrid Analog/Digital Transmission

In LTE/LTE-A network systems and many rapidly-deployable mesh network systems, energy-efficient and delay-limited transmission of analog samples with minimal distortion at the receiving is needed. For such systems, hybrid analog-digital (HAD) is a promising transmission technique. Several papers (see [1] and [2]) have shown that HAD transmission can greatly reduce the decoding latency of the system. In addition, HAD systems have some of the advantages of digital system and some of the advantages of analog system. They can theoretically achieve the Shannon rate-distortion-capacity limit at the designed signal-to-noise ratio. Although HAD is a promising technique for reducing the latency in LTE/LTE-A networks and mesh networks, the use of HAD technique is still face challenges. Because of the fundamental nature of this research, at this early stage it is still not clear how much benefit these techniques will bring with respect to latency in the context of an LTE-like system or rapidly-deployable mesh network systems. Moreover, the application examples of HAD transmission technique in communication systems are very rare. Therefore, many challenging problem have to be solved before this technique can be applied in real communication systems.

In terms of hybrid D/A transmission for LTE/LTE-A network, it is important to demonstrate that a latency-constrained digital feedback protocol for transmission of analog samples can achieve comparable distortion performance to classical feedback-based control approaches without latency constraints. Suggestions for accommodating this type of transmission in and LTE framework will be provided. The more general multi-sensor case will be also considered.

Regarding HAD transmission technology for topology B, we first focus on link level optimization. This includes performance optimization of single-link HAD techniques for different channel models and various source distributions. System level optimization will also be considered. We will also design and optimize the HAD transmission technology for the defined mesh network scenarios. Practical HAD transmission schemes that can reduce latency in topology B will be proposed.

7. Conclusion

Latency is becoming a key issue for network operators seeking solutions to support low latency applications, and in particular realtime machine-to-machine communication and

interactive multiplayer game. Latency is identified as a major factor influencing the behavior and footprint of the application. Currently, LTE can provide on the order of 10ms latency for the E-UTRAN in the ACTIVE state. The core network adds a significant amount of delay depending on the region and the proximity of the server with respect to the access network serving the device. The upcoming emerging application scenarios will definitively require very low access layer latency. A significant latency improvement is possible through advanced access layer techniques as well as careful selection of various parameters and technologies.

8. Acknowledgment

This paper describes work undertaken in the context of the LOLA project - Achieving LOw-LAtency in Wireless Communications (www.ict-lola.eu). The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement n° 248993.

9. References

[1] Skoglund, M., Phamdo, N. and Alajaji, F., "Hybrid Digital-Analog Source-Channel Coding for Bandwidth Compression/Expansion". IEEE Transactions on Information Theory, vol. 52, no. 8, pp. 3757-3763, August 2006.

[2] Mittal, U. and Phamdo, N., "Hybrid digital-analog (HDA) joint source-channel codes for broadcasting and robust communications". IEEE Transactions on Information Theory, vol. 48, no. 5, pp. 1082-1102, May 2002.

[3] I.Tomic, S.Krco, D.Vuckovic, A.Gluhak, P.Navaratnam, “SENSEI traffic impact on mobile wireless networks”, Towards the Future Internet, pp. 257-266, IOS Press, 2010.

[4] 3GPP TR 25.912 V.9.0.0, “Feasibility study for evolved universal terrestrial radio access (UTRA) and universal terrestrial radio access network (UTRAN)”, Rel. 9, Dec 2009.

[5] 3GPP TS 22.368 V10.0.0, “Service requirements for machine-type-communication (MTC)”, March 2010. Available: http://www.3gpp.org/ftp/Specs/archive/22_series/22.368/

[6] M. Claypool, K. Claypool, “Latency and Player Actions in Online Game”, Communication of the ACM, 2006

[7] LOLA consortium, “Deliverable D2.1, D3.2, and D4.1”, FP7 EU LOLA Project, March 2010. Available: http://www.ict-lola.eu/deliverables/