A Comparison Between One-way Delays in Operating HSPA and LTE Networks

Markus Laner*, Philipp Svoboda*, Peter Romirer-Maierhofer‡, Navid Nikaein†, Fabio Ricciato‡§ and Markus Rupp*
*Vienna University of Technology, Austria, Email:{mlaner, psvoboda, mrupp}@nt.tuwien.ac.at
‡Telecommunications Research Center Vienna (FTW), Austria, Email:{romirer, ricciato}@ftw.at
†EURECOM, France, Email:{nnikaein}@eurecom.fr
§University of Salento, Italy

***Abstract*—The detailed understanding of packet delays in modern wireless networks is crucial to optimize applications and protocols. We conducted high precision latency measurements in operational LTE and HSPA networks, deploying a hybrid approach of active probing and withe-box testing. It allowed us to separately assess the one-way delay contributions of the radio access network and the core network for both technologies. The results show that LTE outperforms HSPA in the case of medium to high data rates. However, due to differences in the radio access procedures, the HSPA uplink connection offers lower delay for specific traffic patterns. A comparison between our measurement results and the requirements for delay sensitive applications exhibits that LTE is not (yet) the generally preferable technology. Hence, further optimizations of the LTE scheduling and resource allocation policies are required to fully exhaust all feasible latency improvements.**

I. Introduction

Long Term Evolution (LTE) is one of most recent communication standards. It is specified by the 3rd Generation Partnership Project (3GPP) consortium as an evolution of the 3G standards Wide-band Code Division Multiple Access (WCDMA) and High Speed Packet Access (HSPA) and, as such, its deployment rate is expected to quickly grow in the next few years. The performance targets of LTE compared to HSPA Rel. 6 are [1]: (i) ten times higher peak user throughput, (ii) two to four times higher spectral efficiency and, (iii) two to three times lower latency. Nevertheless, the preceding technology HSPA will further evolve in future releases and, in the short-term, even compete with LTE.

In what follows, we will investigate the last of the above mentioned performance targets of LTE: latency. In the literature *latency*, or equivalently One-Way Delay (OWD), is defined in numerous ways. The 3GPP specifications [2] define two types of latency, *control-plane latency* and *user-plane latency*. Whereas control-plane latency is the time required by the call-setup procedure, user-plane latency describes the one-way transmit time of a data packet at the Internet Protocol (IP) layer from the User Equipment (UE) to the Radio Access Network (RAN) edge-node or vice versa. For the rest of this work we exclusively consider *user-plane latency*, in the sense of OWD. Note that the 3GPP definition [2] leaves room to some ambiguity in case of fragmented IP packets. Since we are interested in testing also with large packet sizes, we adopt a definition of latency that suits well our purposes while still being compliant with [2].

In this work we aim to verify the above mentioned latency target, by measuring LTE and HSPA latencies in the same setting in order to perform a direct comparison. We adopt a measurement methodology that is an hybrid between active probing and passive monitoring: we inject probe packets into the network and capture it at intermediate interfaces at the edges of the Core Network (CN). In this way, beside comparing OWDs separately for the uplink and downlink directions, we are also able to split the total delay into its RAN and CN components.

The measurements are performed in live operational networks, one LTE and one HSPA Rel. 8 network from the same operator. Special care is taken to ensure a fair comparison between both technologies: we transmit packets to both networks, following the same sending patterns, with the same modem connected to base stations located at the same site.

Performing this work required to solve a number of practical and conceptual issues. On the practical side, we had to synchronize and coordinate the devices involved in the active (for sending / receiving probe packets to / from the network) and passive (for capturing packets inside the network) sides of the measurement setup. In order to identify an IP packet captured at different devices and match the corresponding timestamps, each registered packet must be decoded up to the IP protocol-layer before the respective payload can be extracted and compared. Since the protocol stack at various interfaces is composed differently (especially for LTE and HSPA), every monitoring device has to perform a dedicated decoding operation before reporting the result to a centralized device for comparison.

The main findings of this work are: (i) In downlink LTE displays roughly half of the OWD compared to HSPA. For LTE the latency decreases slightly with increasing the data rate, while for HSPA it stays constant. This leads to similar performance at low data rates, 4 kByte/s or lower. (ii) In the uplink the OWD of HSPA depends strongly on the data rate and packet size: at low rates with large packet sizes the uplink OWD of HSPA is higher than LTE, while for all other setting it is sensibly lower. These differences must be accounted to the different uplink resource allocation strategies between the

two technologies. (iii) Considering a fixed average data rate, the LTE latency is insensitive to packet size. Instead, for HSPA there is a strong positive correlation between packet size and latency. Finally, we note that LTE provides more freedom in configuring the network than HSPA, hence allowing for better tuning the *throughput-latency trade-off*. Consequently, global peak-throughput and latency can be jointly optimized.

II. Related Work

Definitions for latency in LTE networks can be found in [2], where a distinction between *user-plane* and *control-plane* latencies is made. Because of the general character of these definitions and the structural similarities of LTE and HSPA, the definitions are further applicable to the second without modifications. Moreover, there are more detailed latency definitions and estimations by 3GPP, which where discussed at the *TSG-RAN WG2 meeting #53*. A list of the discussed documents is given in the respective minutes [3, Sec. 11.3.2].

Latency assessments available in literature are commonly based on estimations. Pure LTE delay estimations can be found in 3GPP documents [3], in early scientific publications [4] and in industrial reports [5], [6].

Measurement-aided estimation techniques for OWD are very popular. One of the simplest techniques is based on the *ping* program, which measures the Round-Trip Time (RTT) from the client to a remote server in the Internet by sending *ICMP echo requests* to the server. Some examples with focus on LTE and former 3GPP standards are found in [1], [7]–[12]. The OWD is thereby derived by assuming symmetric links and simply halving the RTT. However, the assumption of symmetric links does not hold true for mobile cellular systems, and our measurements confirm that the two directions display very different latencies.

True OWD measurements require the capturing and timestamping of data packets at both ends of the connection link. This most often involves distributed and synchronized measurement nodes [13]–[16]. The authors of [13] present a complete measurement tool for evaluation of communication links in terms of various metrics. Those are among others: jitter, packet loss, throughput and OWD. The timing is thereby provided by Global Positioning System (GPS) receivers and a custom software solution, the respective accuracy is estimated to below 100 μs. The authors of [14] assessed the performance of different HSPA networks in terms of OWD. Their measurement devices are similar to those deployed within this work, whereas the timestamping accuracy is estimated to below 100 ns [17, pp. 97-98]. The difference to our work is the traffic type generated for active probing, for which we do not use the *ping* program.

Latency analyses with direct access to mobile network components are rare in literature. Some examples are [18]–[20], which perform passive large-scale RTT measurements by monitoring the TCP-handshakes of mobile users, and [21]–[23] where OWD within the CN for a UMTS network were provided. The passive monitoring system of those measurement campaigns has been partly reused for the present work.

Our present measurement approach is *distributed OWD assessment by white-box testing with active probing*. We have already presented this approach for HSPA networks in [15], [16]. Those measurements comprise uplink and downlink OWDs in live operational networks, where we focus on determining the delay contributions of single network elements. The measurement setup therein is similar to this work, with appropriate extensions (especially on the passive monitors) for covering also the LTE section.

III. Measurement Setup

We carried out delay-measurements in an HSPA and an LTE network in Vienna, Austria, in the first quarter of 2012. The HSPA network operates in Frequency Division Duplex (FDD) Dual-Cell (DC) mode, resulting in a bandwidth of 10 MHz. For LTE the bandwidth is 20 MHz, operating in FDD as well. Both networks belong to the same mobile operator and are publicly accessible. Hence, both networks were hosting an unknown number of users during our measurements. According to the commercial situation, it is safe to assume that the LTE network was exposed to a considerably lower load than HSPA.

In order to reduce daily and weekly effects we performed three measurement runs for each technology, on three successive days and at different hour. For the following evaluation we consider only one dataset out of those three, and specifically the one for which the network appeared least loaded.

An overview of the measurement scenario is given in Fig. 1. A client computer is initiating a bi-directional data transfer to a remote server in the Internet. Probe traffic consisted of two independent User Datagram Protocol (UDP) flows in the two directions, with packet sizes and inter-departure times generated randomly as described later in Sec. IV. Note that we tested with UDP datagrams of size up to 5 kBytes, therefore IP fragmentation takes place for large datagrams.

The client is connected via USB to a triple-mode LTE modem, namely, Huawei E392 [24]. Depending on the measurement run, the modem is manually enforced to access only one of the available communication technologies, either HSPA or LTE. The modem was kept fixed during all our measurement runs in direct line-of-sight to the base stations. Further, the antennas of the Base Stations (NodeBs) of both technologies where located at the same site at a distance of roughly 130 m to the modem. The configurations of the networks where both according to 3GPP Rel. 8, what yields a certain similarity in the network structure. The RAN consists of (i) the NodeBs and the Radio Network Controller (RNC) in HSPA and (ii) the Evolved Base Station (eNodeB) in LTE, which is handling all functionalities related to the radio interface. Furthermore, the CN, denoted as System Architecture Evolution (SAE) in LTE, consists of (i) the Gateway GPRS Support Node (GGSN) in HSPA, which is the only element on the data path since Rel. 7 and (ii) the SAE Gateway (SAE-GW) in LTE.

The OWD measurements have been performed by recording and accurately timestamping IP packets at the different interfaces between the above mentioned network components. This is indicated by the labels *Probe* in Fig. 1. In order to capture

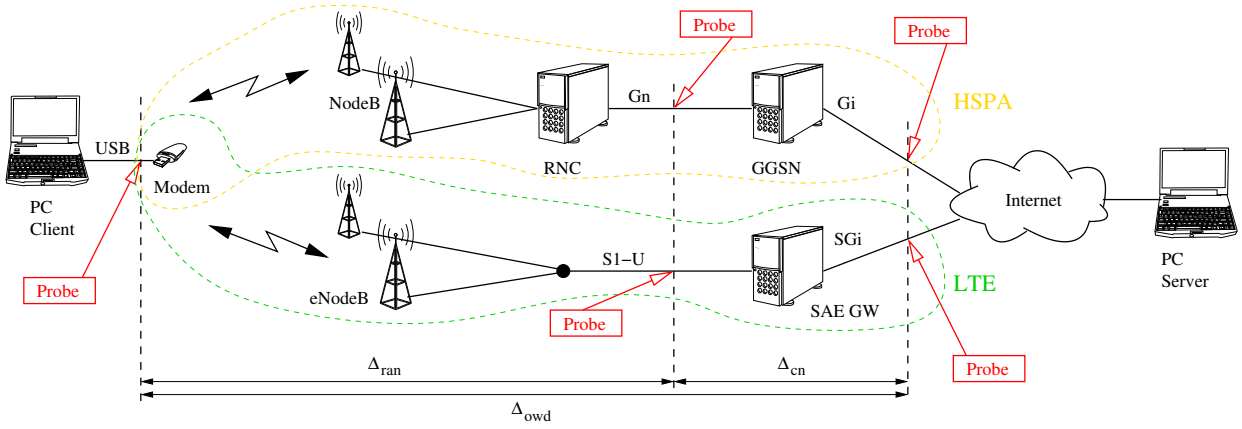

Fig. 1. Measurement setup. Bi-directional data traffic is generated between client and server. The upper part shows the HSPA network, the lower part the LTE network. The dashed lines define radio access and core networks. The boxes labeled *Probe* indicate the wiretapped links from which the measurements are obtained.

packets at the five separate points in the mobile networks, we deployed two types of measurement devices, described in Sec. III-A and III-B.

The OWD for each IP packet was calculated by comparing the timestamps recorded at three different interfaces along the data path. For LTE we compared t_{USB}, $t_{\text{S1-U}}$ and t_{SGi}, whereas for HSPA we checked t_{USB} against t_{Gn} and t_{Gi}. In order to guarantee the timestamps used for evaluation belong to the same packets, we verified that three indicators were equal for every packet at each interface: (i) the IP identification number, (ii) the packet size and, (iii) the first 10 Bytes of the UDP payload. The UDP datagram is mapped to a single IP *datagram*. However, if size of the latter exceeds a certain interface-dependent threshold, called Maximum Transmission Unit (MTU), they are fragmented into smaller IP *fragments*. This results in multiple timestamps $t_{\text{X,i,k}}$ per datagram i and interface X, with index k referring to the kth fragment. Note that both the originally IP datagram and the smaller IP fragments are referred to as IP *packets*. That leaves room to a certain ambiguity in the 3GPP latency specification [2] that refers only to IP *packets*.

Since the end-terminals must necessarily reconstruct the whole datagram before passing it to the upper layer, we choose to focus on datagram latency, as this metric is closer to the user experience. Moreover, it is convenient to adopt a definition for latency that is additive: if the *datagram* is routed over a path consisting of multiple sections, with individual *fragments* being independently forwarded at each node, the total latency should equal the sum of latencies in each section.

In order to fulfill this requirement, we define the datagram latency between two interfaces as the differences between the largest timestamp at each interface, namely, the timestamp of the last fragment associated to the datagram. Formally, the latencies for the ith datagram are calculated as

$$\begin{aligned}\Delta_{\text{ran,i}} &= |\max_{\text{k}}(t_{\text{S1U/Gn,i,k}}) - \max_{\text{f}}(t_{\text{USB,i,k}})| \\ \Delta_{\text{cn,i}} &= |\max_{\text{k}}(t_{\text{SGi/Gi,i,k}}) - \max_{\text{f}}(t_{\text{S1U/Gn,i,k}})| \\ \Delta_{\text{owd,i}} &= |\max_{\text{k}}(t_{\text{SGi/Gi,i,k}}) - \max_{\text{f}}(t_{\text{USB,i,k}})|. \end{aligned} \quad (1)$$

A. Client Monitoring

The Client has two tasks in the measurement setup: (i) it hosts the traffic generator, as described in Sec. IV and (ii) it runs the software for timestamping and recording the transmitted packets. Since we want to avoid any influence of the scheduling of the operating system on the client, we perform the timestamping of the packets in the kernel-domain, using *libpcap* [25]. For the recording of the packets the *tshark* software was deployed [26]. In order to get meaningful timestamps, the client was synchronized to Coordinated Universal

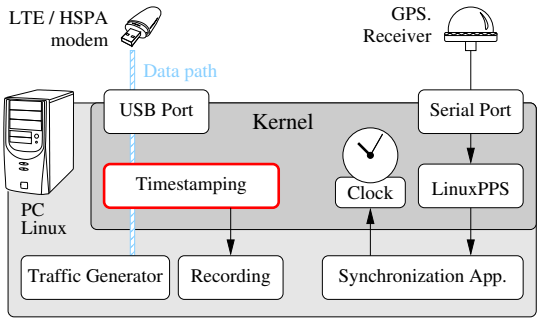

Fig. 2. Detail of the measurement probe at USB. The client PC has the task of traffic generation and synchronized time-stamping.

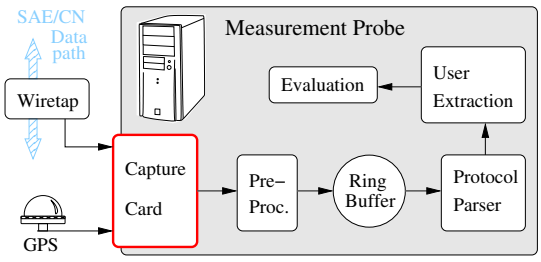

Fig. 3. Detail of the measurement probes within the mobile networks. The probes timestamp captured packets and decode them for identification and packet-matching.

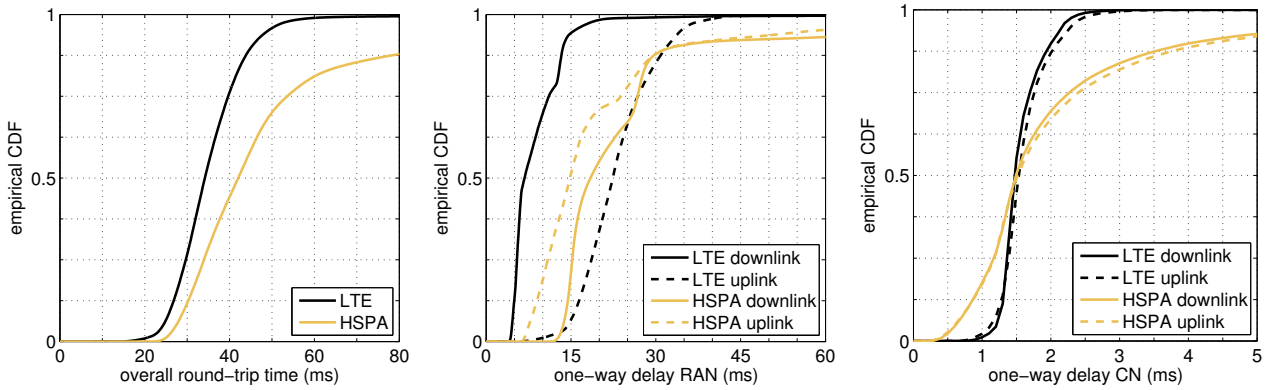

Fig. 4. Latency of different technologies for a broad range of packet sizes and data rates. (left) CDFs of the overall round-trip time for HSPA and LTE. (center) CDFs of the one-way delay caused by the radio access network. (right) CDFs of the one-way delay caused by the core network.

Time (UTC). We choose *LinuxPPS* framework [27] for this task, which enables a synchronization accuracy of approx. 10 μs [28]. Fig. 2 gives a schematic overview on the described setup.

B. Core Monitoring

The capturing of packets at connections within the providers SAE/CN (i.e., at each link except the USB link), was performed by dedicated measurement equipment developed within the DARWIN project [29], [30]. All links are traced by measurement probes which deploy wiretaps with high-rate data acquisition cards [31] and GPS synchronization with sub microsecond accuracy [17, pp. 97-98]. Because of the high data aggregation at the SAE/CN interfaces, the main problem with tracing consists of the extraction of the packets dedicated to the user of interest. This requires measurement nodes with high-speed protocol parsing capabilities and the possibility to correlate information from different protocol layers. The differences between the four probes deployed at different interfaces (Gn, Gi, S1-U, SGi) are mainly in software. Since each of the assessed interfaces inherits a different protocol stack, each monitoring probe needs respective parsers. Fig. 3 shows the schematic setup.

IV. Data Traffic Generation

3G/4G mobile networks are *stateful* systems: due to *Radio Resource Management* mechanisms (e.g., adaptive channel assignments) the delay experienced by each generic packet depends heavily on the past history, namely, on the timing and size of the preceding packets transmitted and received by the same mobile terminal. Therefore, different traffic patterns will unavoidably result in different OWD distributions. Moreover, while in fixed-capacity wired networks one can expect that increasing the data rate will result in larger packet delays (due to queuing), this is not a priori true in mobile networks, wherein radio bandwidth is assigned dynamically to each terminal, depending on its actual traffic demand. Therefore, higher traffic rates lead to smaller delays in specific cases. This is a fundamental differences between wired and mobile networks which invalidates the applicability of many bandwidth estimation methods to mobile networks. Based on such considerations, it should be clear that delay statistics obtained with a specific traffic pattern must be interpreted with caution and cannot be considered as exactly representative of the delay experienced by an application with a different traffic pattern.

Since our goal here is to benchmark two networks, LTE and HSPA, the probe pattern must be designed in order to avoid any possible bias. To avoid synchronization effects (LTE and HSPA networks are synchronous) and guarantee the PASTA property [32], [33], packet timings need to be randomized as proposed in RCF 2330 [8], [34]. Probe packets are generated with random size and inter-arrival time.

The traffic generator deployed for our measurements is a custom user-space application. It produces two independent UDP streams for uplink and downlink, both at port 3000. Since we aim to investigate the influence of the data rate on the OWD, the probe stream is organized into chunks of fixed data rate R. Each chunk has a duration of 150 s, followed by an idle period of 10 s, that forces the scheduling processes at the NodeB to return to a common state. Within each chunk the packet (datagram) sizes at UDP layer s_i are random, uniformly distributed between 10 and 5000 Byte. The high upper limit shall enable to observe IP fragmentation effects. The packet inter-arrival times d_i are calculated for each packet i as $s_i = R \cdot d_i$. Consequently, the distribution of inter-arrival times is also uniform, with limits depending on the data rate R. Additionally, for all chunks we impose a lower limit of 1 ms and an upper limit of 1 s. We prefer uniform distributions to the Poisson [32] or Gamma distributions [33], in order to obtain a constant number of samples for all packet sizes, which further allows for omitting normalization steps in the statistical evaluation.

V. Results

A. Round-trip Time and Individual One-way Delays

Fig. 4 compares the Cumulative Distribution Functions (CDFs) of LTE and HSPA Rel. 8 latencies.

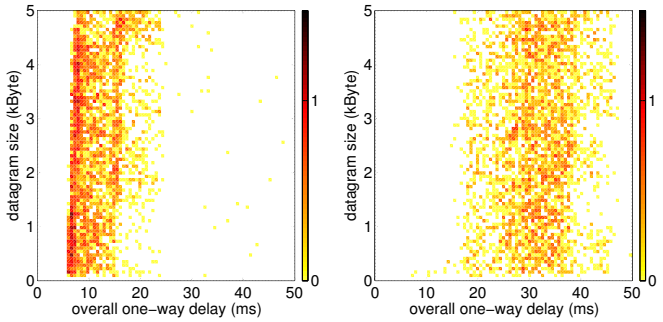

Fig. 5. Scatter-plot of LTE latency vs. datagram size at $R = 64$ kByte/s: downlink (left) and uplink (right). Number of packets in normal logarithmic color code (yellow ~ 1, dark red ~ 10). No correlation between one-way delay and datagram size.

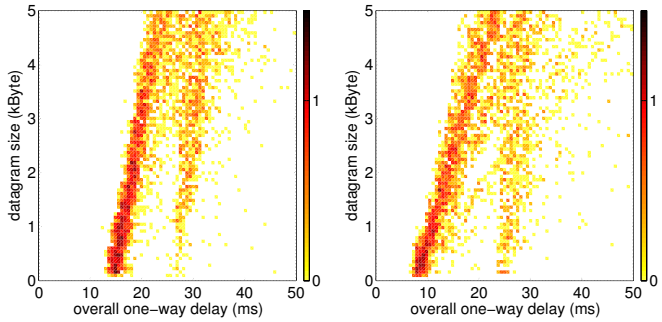

Fig. 6. Scatter-plot of HSPA latency vs. packet size at $R = 64$ kByte/s: downlink (left) and uplink (right). Strong correlation (diagonal components) between packet size and delay, especially for uplink. In downlink the correlations are vanishing for $R > 256$ kByte/s.

RTT: In the leftmost graph the overall RTTs are compared. Since we did no RTT measurements, the respective values are synthesized from the uplink and downlink OWDs taken as independent random variables, namely, from the convolution of the respective Probability Density Functions (PDFs). It is observable, that the latency reduction brought by LTE is small, namely, 36 ms compared to 42 ms (or 14%) by consulting the median. However, comparing the minimum delay, we note that LTE is able to perform much better than HSPA (11.9 ms vs. 18.3 ms or 35% reduction). We refrain from consulting mean or higher percentiles, because of the highly different load within both networks that causes a long-tail in the HSPA latency distribution.

OWD in RAN: The delay contribution of the RAN is presented in Fig. 4 (center). We find that the downlink OWD in LTE is significantly lower than in HSPA. For the uplink the situation is reversed. The reason is the difference in resource allocation: whereas in LTE the mobile has to request time slots for each packet, the corresponding scheduler in HSPA allocates code-power for the average data rate, so that there is no extra negotiation cycle between UE and NodeB. The *semi-persistent scheduling* mode for LTE improves this situation, since it enables to permanently allocate radio resources to a mobile station. However, this leads to the further problem of choosing the amount of reserved resources for each mobile.

OWD in SAE/CN: The delay contributions through the SAE-GW (for LTE traffic) and the GGSN (for HSPA traffic) are shown in Fig. 4 (right). All different OWDs are minor compared to the overall RTT (note the different scale on abscissa). As noted already in [16], the CDFs for uplink and downlink OWD match very well for both technologies. Moreover the median OWD corresponds exactly to the values indicated by 3GPP [3].

B. One-way Delay and Packet Size

Correlations between latency and packet size are reported in various works [8], [10]. The respective assessment within our study is done by the latency vs. packet size scatter-plots in Fig. 5 and Fig. 6 for LTE and HSPA respectively. The color code in normal logarithmic scale reveals the number of packets per bin. The plots show all packets of one chunk at rate 64 kByte/s. The plots for other chunks at different rates are very similar and are omitted for sake of space

From Fig. 5 we observe that for LTE the OWD in both directions are insensitive to packet size. This behavior also applies for *all* other data rates (graphs omitted) in both directions. Instead, Fig. 6 shows that for HSPA the OWD are positively correlated to the packet size (skewed components). The correlation is stronger in uplink. In downlink, the correlation decreases with increasing data rates and vanishes above 256 kByte/s. We already reported on this behavior in previous work [16].

The difference in latency-size correlations between the two technologies can be accounted to the different resource allocation policies. In LTE, as well as in HSPA-downlink, the scheduling relies on the immediate request of resources. As reaction on the request a proper amount of resources is allocated. For the downlink direction this procedure takes place in the NodeB, without causing considerable delay. For LTE-uplink, on the other hand, a handshake procedure is introduced, consisting of a *scheduling request* message from the UE and a *scheduling grant* from the NodeB [1, pp. 97]. This handshake requires twice a communication over the air interface, causing notable delay. In HSPA-uplink instead, resources are steadily allocated to the mobile station in terms of code power. If the overall rate changes, the mobile station requests additional resources by swapping the *happy bit* [35, pp. 85]. This explains the good performance for small packets, as well as the extra latency for big packets.

C. One-way Delay and Data Rate

In oder to understand the impact of data rate we plot in Fig. 7 the median OWD obtained at different rates. The median allows for a more fair comparison even if the two networks operate at different loads. We focus here on packet size below 1500 Bytes. For HSPA we plot two distinct curves per direction, one for small packets (size below 250 Bytes, lower curves) and one for large packets (size between 1250 and 1500 Bytes, upper curve).

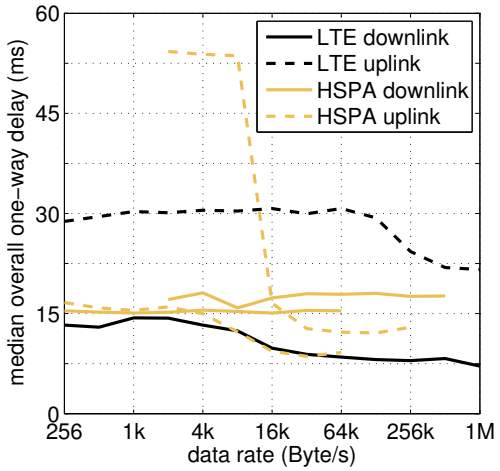

Fig. 7. Median one-way delay over data rate for different technologies. LTE: all packets smaller than 1500 Bytes. HSPA: two curves are shown per each direction, one for packets smaller 250 Bytes (lower curves), the other for packet sizes between 1250 and 1500 Bytes (upper curves). The reason is the observed correlation of packet size and delay in HSPA (see Fig. 6). Generally OWD decreases with increasing data rate. HSPA outperforms LTE for small packets and/or low rates.

The general trend in Fig. 7 shows that the delay is decreasing for an increasing data rate. In the uplink direction this effect is more distinct. As already asserted in Sec. V-B it is again visible in Fig. 7 that the dependence of the HSPA-latency on the packet size is stronger in uplink than in downlink. Especially for uplink data rates lower than 16 kByte/s we observe this dependence to cause a change in delay of a factor of three. Further, for small data rates and low variations of the packet sizes, HSPA may show lower uplink OWD than LTE. The reason being the different uplink scheduling policies in both technologies described above. This explains the small gain in RTT of LTE compared to HSPA, which we observed in Fig. 4.

VI. Interpretation of One-way Delay Analysis

In the previous section we analyzed the technical aspects of the OWD figures for LTE and HSPA. Now we map these results onto the applications and traffic types of the transport network, in order to understand if LTE is the preferred access network for all applications. As noted earlier in Sec. IV, the delay statistics obtained with our probe traffic generator cannot be taken as exactly representative of the delay experienced by real applications. Nevertheless, based on the measurements above, we can derive indications and qualitative predictions about the delay performance that can be expected by different classes of applications in each technology.

In the following we will focus on a set of delay sensitive applications. Note that applications like *file-downloads* or *web-browsing* are not considered here, since the user-experience for those cases is only affected by RTTs bigger than 1 s [36]. The considered traffic types are

- Online Gaming,
- Machine-to-Machine Communication (M2M) and
- Voice over IP (VoIP)

Application	LTE (up/down)	HSPA (up/down)
Online Gaming	(31 / 13) ms	(12 / 17) ms
M2M	(30 / 10) ms	(10 / 16) ms
VoIP	(30 / 15) ms	(35 / 16) ms

TABLE I
Delay per Application Type - Suitability

A. Properties of the Different Traffic Types

The first group of applications, Online Gaming, is characterized by small constant sized packets with random inter-arrival times in uplink, and variable sized packets with constant arrival time in downlink. The data rate is between 1 kByte/s and 5 kByte/s for most cases [37]. The delay required for good Quality of Service (QoS) lies below 50 ms [36].

The second group of applications is M2M traffic. Analysis of emerging M2M application scenarios such as smart metering/monitoring, e-health, and e-vehicle has reveled that in majority of cases, packets are short and small in number and extremely low duty-cycle, which from a system throughput perspective represents a vanishing data rate [37]. In addition, such packets are more uplink-dominant and follow two traffic patterns: (i) periodic non-realtime packets such as keep alive or update messages and, (ii) event-driven realtime packets such as alarm notifications. The latter case requires a fast reaction time from an overall system latency point of view. The delay requirements for proper functionality of future applications is debatable, still, M2M applications may appear which demand RTTs below 25 ms [36] (e.g., pipeline alarm-sensors, control-loops with wireless feedback).

The last group of applications, VoIP traffic, shows constant small-sized packets with a constant inter-arrival time, hence, a constant data rate. The vast amount of different VoIP applications makes the firm characterization of traffic of this type rather difficult. Data rates may reach from 4 kByte/s up to 200 kByte/s. Further, RTTs of up to 200 ms are sufficient for excellent user experience.

B. Best Technology for each Application

To answer the question which technology is best suited for these traffic patterns we examine the information given in Sec. V. The interactive traffic of Online Games benefits from the lower delay in the downlink of LTE. A drawback is the increased uplink delay. The sporadic nature of M2M traffic patterns and the main traffic flowing in uplink (most M2M nodes are assumed to be sensors), leads to the conclusion that the latency performance of HSPA will be better in this case. In VoIP both directions shall offer a delay below 100 ms. This can be achieved by both technologies. The higher delay of LTE in certain situations favors HSPA for some applications. The results are summarized in Table I. Therein, *green* signifies high user satisfaction, whereas *red* indicates that the QoS may be impaired. The coloring is based on the results obtained in [36], where a detailed analysis for QoS parameters is given for these applications.

The fact that there is no circuit-switched connection mode in LTE, motivated the introduction of *semi-persistent scheduling*, which, beside of reducing the scheduling signalization overhead, strongly reduces the latency. It was mainly designed for VoIP traffic, but due to the flexibility offered by LTE also other applications may profit from it. Correspondingly deployed, this scheduling mode enables the harvesting of the strengths of LTE latency performance (e.g., minimum measured delay of roughly 5 ms in downlink and 6 ms in uplink).

VII. Summary and Conclusions

In this work the latency-performance of LTE and HSPA networks are compared. We obtained delay figures from public networks following an hybrid measurement methodology that combines active probing with passive monitoring (or white-box testing). The latter enables high-precision analyses of partial OWD components for the RAN and CN network sections. The active traffic generation allows to obtain the delay figures for well defined data rates and packet sizes.

The results show that, as expected, the RTT of LTE is lower (median 33 ms) compared to HSPA (median 42 ms) by roughly 14%. In general all delay values tend to decrease for higher data rates, which is a distinguishing difference between 3G/4G networks and fixed-capacity wired networks. In downlink direction LTE clearly outperforms HSPA, with a median of 8 ms compared to 18 ms. In uplink connection, on the other hand, HSPA shows lower latency (median 15 ms) than LTE (median 24 ms). This is caused by a different resource allocation procedure for both technologies. However, measurements suggest that if the *semi-persistent scheduling* mode will be broadly deployed for latency sensitive applications, the LTE uplink latency can be reduced to less than 10 ms. In Table ,I we map the latency requirements of different applications onto the expected delay of LTE and HSPA.

Our results indicate that there is room for further research on LTE scheduling policies friendly to latency-sensitive applications. One possible approach is to adjust the *latency-throughput trade-off* on a per-application basis.

Acknowledgments

The authors would like to thank *A1 AG* for providing parts of the technical equipment. The work was supported by the European Union within the *EU-FP7 LoLa* project and by the Austrian Government within the *COMET program* in the *DARWIN3* project.

References

[1] H. Holma and A. Toskala, *LTE for UMTS, OFDMA and SC-FDMA Based Radio Access*. Wiley, 2009.

[2] 3GPP. TS 25.913, Requirements for Evolved UTRA and Evolved UTRAN, http://www.3gpp.org/.

[3] ——. R2-062314, Minutes of the 53rd TSG-RAN WG2 meeting, Shanghai, China, http://www.3gpp.org/ftp/tsg_ran/WG2_RL2/TSGR2_53/Report/.

[4] T. Blajic, D. Nogulic, and M. Druzijanic, "Latency Improvements in 3G Long Term Evolution," in *MIPRO'07, Opatija, Croatia*, 2007.

[5] D. Singhal, M. Kunapareddy, and V. Chetlapalli, "LTE-Advanced: Latency Analysis for IMT-A Evaluation," Tech Mahindra, Tech. Rep., 2010.

[6] S. Mohan, R. Kapoor, and B. Mohanty, "Latency in HSPA Data Networks," Qualcomm, Tech. Rep., 2011.

[7] J. Liu, P. Tapia, P. Kwok, and Y. Karimli, "Performance and Capacity of HSUPA in Lab Environment," in *VTC Spring 2008, Singapore*, 2008.

[8] J. Fabini, L. Wallentin, and P. Reichl, "The importance of being really random: methodological aspects of IP-layer 2G and 3G network delay assessment," in *ICC'09, Dresden, Germany*, 2009.

[9] J. Robson, "The LTE/SAE Trail Initiative: Taking LTE/SAE from Specification to Rollout," *IEEE Communications Magazine*, vol. 47 (4), pp. 82–88, 2009.

[10] C. Serrano, B. Garriga, J. Velasco, J. Urbano, S. Tenorio, and M. Sierra, "Latency in Broad-Band Mobile Networks," in *VTC Spring 2009, Barcelona, Spain*, 2009.

[11] M. Wylie-Green and T. Svensson, "Throughput, Capacity, Handover and Latency Performance in a 3GPP LTE FDD Field Trial," in *Globecom'10, Miami, Florida*, 2010.

[12] L. Schumacher, G. Gomand, and G. Toma, "Performance Evaluation of Indoor Internet Access over a Test LTE Mini-Network," in *WPMC'11, Brest, France*, 2011.

[13] J. Prokkola, M. Hanski, M. Jurvansuu, and M. Immonen, "Measuring WCDMA and HSDPA Delay Characteristics with QoSMeT," in *ICC'07, Glasgow, Scotland*, 2007.

[14] P. Arlos and M. Fiedler, "Influence of the Packet Size on the One-Way Delay in 3G Networks," in *PAM'10, Zurich, Switzerland*, 2010.

[15] M. Laner, P. Svoboda, and M. Rupp, "Dissecting 3G Uplink Delay by Measuring in an Operational HSPA Network," in *PAM'11, Atlanta, Georgia*, 2011.

[16] ——, "Latency Analysis of 3G Network Components," in *EW'12, Poznan, Poland*, 2012.

[17] S. Donnelly, "High Precision Timeing in Passive Measurements of Data Networks," Ph.D. dissertation, University of Waikato, Hamilton, New Zealand, 2002.

[18] F. Vacirca, F. Ricciato, and R. Pilz, "Large-Scale RTT Measurements from an Operational UMTS/GPRS Network," in *WICON'05, Budapest, Hungary*, 2005.

[19] P. Romirer *et al.*, "Network-Wide Measurements of TCP RTT in 3G," in *TMA'09, Aachen, Germany*, 2009.

[20] P. Romirer, A. Coluccia, and T. Witek, "On the Use of TCP Passive Measurements for Anomaly Detection: A Case Study from an Operational 3G Network," in *TMA'10, Zurich, Switzerland*, 2010.

[21] F. Ricciato, E. Hasenleithner, and P. Romirer-Maierhofer, "Traffic analysis at short time-scales: an empirical case study from a 3G cellular network," *IEEE Trans. on Network and Service Management*, vol. 5, pp. 11–21, 2008.

[22] P. Romirer, F. Ricciato, and A. Coluccia, "Explorative analysis of one-way delays in a mobile 3G network," in *LANMAN'08, Cluj-Napoca, Romania*, 2008.

[23] ——, "Towards Anomaly Detection in One-Way Delay Measurements for 3G Mobile Networks: A Preliminary Study," in *IEEE/IFIP IPOM'08, Samos Island, Greece*, 2008.

[24] Huawei Device, http://www.huaweidevice.com/.

[25] libpcap - library for network traffic capture, http://www.tcpdump.org/.

[26] Wireshark - network protocol analyzer, http://www.wireshark.org/.

[27] LinuxPPS, http://wiki.enneenne.com/index.php/LinuxPPS_support.

[28] M. Laner, S. Caban, P. Svoboda, and M. Rupp, "Time Synchronization Performance of Desktop Computers," in *ISPCS'11, Munich*, 2011.

[29] Darwin Project, http://userver.ftw.at/∼ricciato/darwin/.

[30] F. Ricciato, "Traffic monitoring and analysis for the optimization of a 3G network," *IEEE Wireless Communications*, vol. 13, pp. 42–49, 2006.

[31] Endace DAG, http://www.endace.com/.

[32] R. Wolff, "Poisson Arrivals See Time Averages," *Operations Research*, vol. 30 (2), pp. 223–231, 1982.

[33] F. Baccelli *et al.*, "On Optimal Probing for Delay and Loss Measurement," in *IMC'07, San Diego, California*, 2007.

[34] V. Paxson *et al.* (1998) Framework for IP Performance Metrics, http://www.ietf.org/rfc/rfc2330.txt.

[35] H. Holma and A. Toskala, *HSDPA/HSUPA for UMTS, High Speed Radio Access for Mobile Communications*. Wiley, 2006.

[36] LoLa Consortium, "D3.6, QoS Metrics for M2M and Online Gaming, http://www.ict-lola.eu/," EU-FP7, Tech. Rep., 2012.

[37] ——, "D3.5, Traffic Models for M2M and Online Gaming Network Traffic, http://www.ict-lola.eu/," EU-FP7, Tech. Rep., 2012.