Dissecting 3G Uplink Delay by Measuring in an Operational HSPA Network

Markus Laner[1], Philipp Svoboda[1], Eduard Hasenleithner[2], and Markus Rupp[1]

[1] Vienna University of Technology, Austria
mlaner@nt.tuwien.ac.at
[2] Telecommunications Research Center Vienna (ftw), Austria

Abstract. Users expect mobile Internet access via 3G technologies to be comparable to wired access in terms of throughput and latency. HSPA achieves this for throughput, whereas delay is significantly higher.

In this paper we measure the overall latency introduced by HSUPA and accurately dissect it into contributions of USB-modem (UE), base station (NodeB) and network controller (RNC). We achieve this by combining traces recorded at each interface along the data-path of a public operational UMTS network. The actively generated sample traffic covers real-time applications.

Results show the delay to be strongly dependent on the packet size, with random components depending on synchronization issues. We provide models for latency of single network entities as well as accumulated delay. These findings allow to identify optimum settings in terms of low latency, both for application and network parameters.

1 Introduction

In the past few years the number of mobile devices accessing Internet via 3^{rd} Generation (3G) technologies experienced a significant grow. Novel gadgets such as smartphones and netbooks captured a new market, providing Internet access paired with high mobility. Their users expect a connection quality comparable to wired Internet access in terms of throughput and delay. In contrast to their wired counterpart mobile broadband connections have to deal with varying channel conditions depending on a manifold of parameters such as user position, mobility and total number of users in a cell. This causes challenges in hiding limitations of the access technology from the end-application and user.

The state of the art (2010) radio access technologies are High Speed Downlink Packet Access (HSDPA) and High Speed Uplink Packet Access (HSUPA), specified in the 3^{rd} Generation Partnership Project (3GPP). These technologies allow for throughput comparable to wired access, whereas the access delay is still significantly higher. Although improved compared to former releases [1], HSUPA introduces high latency. The reason being the wireless channel as communication resource shared among unsynchronized users and the master-slave hierarchy in 3G networks, meaning the Base Station (NodeB) has to grant access to the User

N. Spring and G. Riley (Eds.): PAM 2011, LNCS 6579, pp. 52–61, 2011.

Equipment (UE) before data can be send. Hence, realtime applications claiming very low latency encounter difficulties when connected via 3G networks. Such realtime applications may be online games or machine-to-machine communication [2]. Application designers can exploit knowledge about delay characteristics of mobile wireless connections to improve user experience. On the other hand, networks can be optimized in terms of latency, given precise information about its origin. Having reached wired data rates, reduction of delay is one of the main goals for next generation wireless networks.

This work investigates the overall uplink One-Way Delay (OWD), Δ, introduced by an operational HSUPA network and analyses the exact delay contribution of every single network component. We confine ourselves to measure OWD because the up and downlink are strongly asymmetric, hence, Round-Trip Time (RTT) measurements have weak significance. Furthermore, we assess latency of the 3G network only, since it constitutes the first *hop* in terms of packet communication. Data packets have been traced and accurately timestamped on each communication link, from the destination PC throughout the UMTS Terrestrial Radio Access Network (UTRAN) up to the Internet gateway. Since each packet is subject to changes in protocols and size, we particularly monitor Internet Protocol (IP) packets, for which the mobile network is transparent. We pay special attention to the packet size, which has strong influence on the OWD.

To the best of our knowledge this is the first work reporting accurate OWD measurements from a HSUPA network, providing latency statistics of each network component. In [3] the authors performed end-to-end measurements of OWD with high timestamping accuracy, however, without intermediate measurement points. They give results for three different network operators. Their traffic generation method differs significantly from ours. The authors of [4] and [5] provide OWD measurements with low timestamping accuracy from multiple network operators. They use *ICMP ping* messages as measured data traffic, in order to highlight the importance of the right data generation method, which has to be *RCF 2330* [6] compliant in their opinion. RTT measurements from a HSUPA testbed are presented in [7], where data was generated by the *ping* program. In [8] large-scale RTT measurements from a Wide-band Code Division Multiple Access (WCDMA) network are presented, resulting from captured Transmission Control Protocol (TCP) acknowledgement packets. Parts of the presented measurement setup have been reused for this work. Furthermore, possible reasons for variability in delay in wireless networks are highlighted, which do mostly apply for HSUPA as well, e.g. radio channel conditions or scheduling and channel assignment. Finally, the authors of [9] investigate OWD introduced by the Serving GPRS Support Node (SGSN), a 3G network component. Although reusing parts of their measurement setup, results cannot be compared because 3GPP specifies that from Rel. 7 on data traffic bypasses the SGSN.

This paper is structured as follows. Section 2 explains the measurement setup in detail. The results are presented in Section 3 and analyzed in detail. We conclude with Section 4, giving an outlook on future networks.

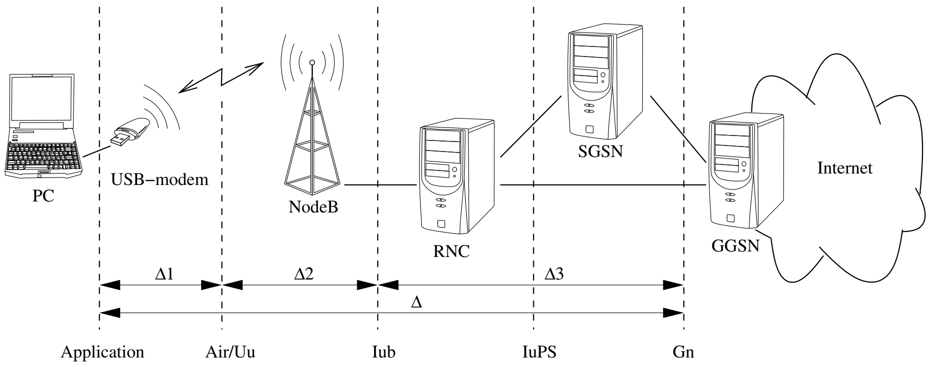


Fig. 1. The UMTS network and its components, 3GPP Release 7

2 Measurement Setup

The measurements were carried out in the operational Universal Mobile Telecommunication System (UMTS) network, of one of the biggest operators in Austria, EU. An overview of the data path of this network is given in Fig. 1. The dashed lines indicate the names of the different interfaces between network components. $\Delta 1$ to $\Delta 3$ indicate the delay contributions of the single elements, Δ the accumulated delay. In the following the components are explained briefly.

- *PC.* The computer on which the end-application is running and application interface traces are captured.
- *USB-modem.* The USB-modem used for measurements is manufactured by Option [10] and equipped with Rel. 7 HSUPA functionality.
- *NodeB.* The Base Station (NodeB) receives and decodes the packets. For controllable measurement conditions an indoor NodeB was chosen.
- *RNC.* The Radio Network Controller (RNC) is the controlling entity in the UTRAN. It coordinates multiple NodeBs. It handles tasks such as ciphering, soft-handover and radio connection manipulations.
- *SGSN.* The Serving GPRS Support Node (SGSN) controls the radio connection and handles mobility issues. Since Rel. 7 it is not part of the data path any more.
- *GGSN.* The Gateway GPRS Support Node (GGSN) is the gateway to the Internet. It sends plain IP-packets towards their destination.

All interfaces shown in Fig. 1, except IuPS, were traced in order to carry out delay measurements of each separate network component. The exact methodology is explained in Section 2.2 for each interface separately. The reason for not tracing the IuPS interface is the *direct tunneling* feature taking effect in Rel. 7. This feature allows the SGSN to remove itself from the data path. Consequently, the expected delay between IuPS and Gn interface is negligible and not considered further.

2.1 Traffic Generation

The traffic patterns sent over the network in order to measure latency were generated actively. According to the proposals in IP Performance Metrics (IPPM) RFC2330 [6], they consist of packets with random size and random-inter arrival time. The importance of the right choice in traffic patterns is highlighted in [5], where the authors reason that invariant traffic generation models such as used by the *ping* command are not adequate for latency measurements in 3G networks. We chose User Datagram Protocol (UDP)-packets for transmission, whereas we allow for large packets up to 10 kByte. This approach is unusual for network measurements, because big packets are segmented into smaller packets of maximum Packet Data Unit (PDU) size. However, the 3G network is transparent for IP packets and interprets segments just as extra payload. Furthermore, such packet sizes are demanded by latency sensitive applications [11], and therefore considered in this work. In order to guarantee the USB-modem is operating in HSUPA mode, we kept the mean data rate above 1 kbit/s. Otherwise the network scheduler would release the HSUPA connection and force the modem to WCDMA Forward Access Channel (FACH) operation, in order to save radio resources. Consequences of such a fallback are observed in [4] and [3], resulting in very high delay values for small packet sizes. In the context of this study these effects are undesired and hence avoided.

2.2 Measurement Devices

OWD measurements require careful consideration of (i) time synchronization of the measurement entities and (ii) accurate packet recognition. In our measurement setup we use Global Positioning System (GPS) receivers for time synchronization, which allow for a precision better than 1 μs. This precision is satisfactory for our purposes, since we plan to achieve a maximum resolution of 100 μs. We use full IP and UDP headers to distinguish between packets at different interfaces. Since the whole 3G network, from UE to GGSN corresponds to one *hop* in terms of IP-networking, both packet headers are not altered during the propagation. In the following sections measurement methods and devices are described.

Gn Interface. As depicted in Fig. 1, the Gn interface connects the GGSN to the rest of the 3G network. We passively monitor this link by means of wiretaps and dedicated tracing hardware, i.e. Endace DAG cards [12] with GPS synchronization. The system has been developed in an earlier project in collaboration

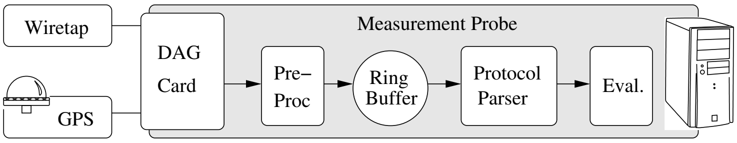


Fig. 2. Measurement setup at the Gn and Iub interfaces

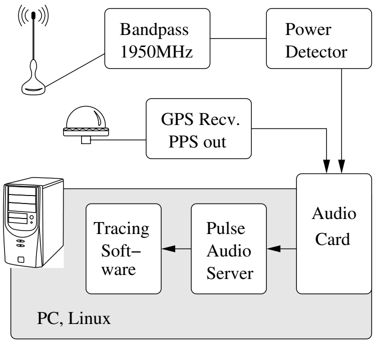


Fig. 3. Transmit power measurement setup (air interface)

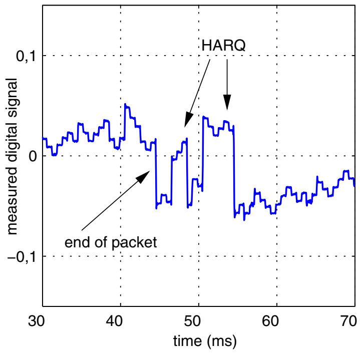


Fig. 4. Measured transmit power of UE (digital domain)

with Telecommunications Research Center Vienna (ftw), see [13] and [14]. An outline of the measurement setup is given in Fig. 2. The timestamping accuracy is specified by the manufacturer with less than 200 ns.

Iub Interface. For data acquisition at the Iub interface the same measurement setup as for Gn has been deployed, see Fig. 2. Tracing at this interface appears particularly challenging because of the complex protocol hierarchy, ciphered payload and *soft handover* [15]. IP packets do not appear in one piece at this interface but split into single Radio Link Control (RLC) frames which are timestamped separately.

Air Interface. Packet sniffing (fully decoding) at the air interface we consider too challenging for our purposes. Instead, we can identify start and end time of single packets by monitoring the transmission power of the UE. This is HSUPA specific, since the NodeB assigns extra transmission power to the UE via Relative Grant Channel (RGCH), in order to transmit data in uplink [1]. This method allows to identify packet transmissions, as long as the inter-arrival time of packets is big enough to guarantee a change in allocated transmission power between packets. Depending on the payload size we varied this time from 10 ms to 100 ms. The measurement setup is depicted in Fig. 3. An antenna with bandpass filter (1920 - 1980 Mhz) and attached power detector [16] is placed nearby the UE. The measured signal is fed into a standard audio device of a PC, with a sampling rate of 44.1 kHz and 16 bit resolution. Figure 4 shows the resulting digital signal. Here we observe the end of a packet transmission (44 ms) with Hybrid Automatic Repeat Request (HARQ) retransmission (46 ms, 50 ms). The small steps result from the Inner Loop Power Control (ILPC) power adjustments. Synchronization is achieved by applying the Pulse Per Second (PPS) output of a GPS receiver [17] at the second audio channel. The timestamping accuracy is limited by the inter-sample time of the audio card (22.7 μs).

Application Interface. We chose the traffic generating application and the application-interface traffic monitoring tool to reside on the same PC. Therefore

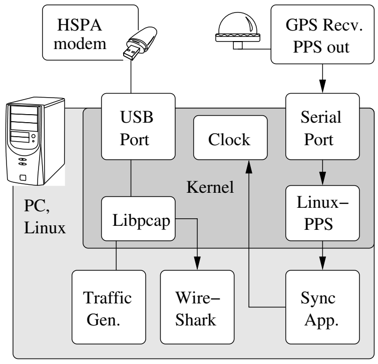


Fig. 5. Application interface measurement setup

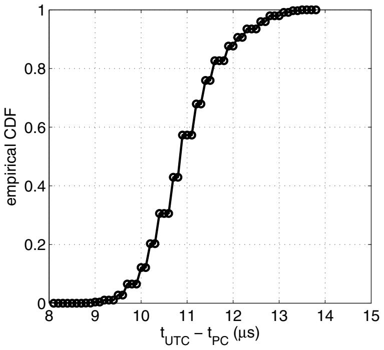


Fig. 6. Synchronization quality

we verify the CPU load to not exceed 20 % during measurements and hence assume the mutual influence of applications to be negligible. Packet capturing was performed by the use of *libpcap* [18] and the *Wireshark* tool, see Fig. 5. In order to achieve correct timestamping of the traffic, we synchronize the software-clock of the PC to Coordinated Universal Time (UTC). We deploy a GPS receiver [17] attached at the serial port and the *LinuxPPS* toolkit [19] to adjust the PC clock, see Fig. 5. The synchronization accuracy was verified with a rubidium oscillator, results yield roughly 10 μs, see Fig. 6.

3 Results

The measurement results presented in the following are obtained from a protected environment. Although, the NodeB to which we established connections is operational and publicly available, it is deployed in an indoor scenario (office) with low cell load and a relatively small number of users. Furthermore, it communicates with the RNC via Asynchronous Transfer Mode (ATM) connection and the Transmission Time Intervals (TTIs) have 10 ms duration. HSUPA also provides 2 ms TTIs for improved latency, hence, the presented results constitute a worst case scenario. The channel conditions were stationary and the data rate was constant in the long run. As pointed out in [8], the deployment scenario strongly influences OWD. We publicly advertise a sample data set [20], enabling reproduction of the following results.

3.1 Single Components

In the following we provide delay measurement results focusing on the single network components, named $\Delta 1$ to $\Delta 3$ in Fig. 1. This information allows to identify main sources of latency and to detect network settings which are improvable in terms of delay.

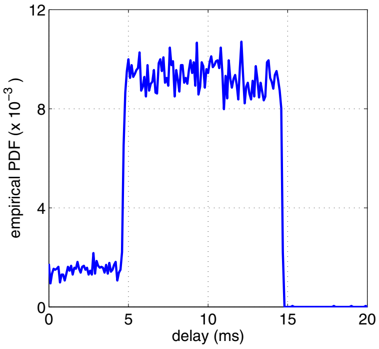


Fig. 7. Delay $\Delta1$, empirical PDF

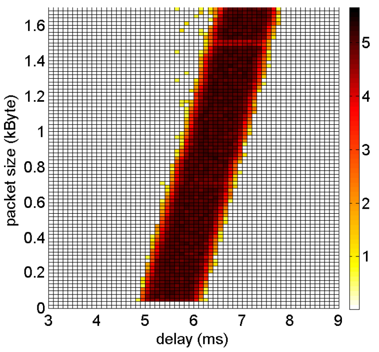


Fig. 8. $\Delta2$ over size, log. histogram

UE. The latency contribution of the USB-modem, $\Delta1$ (see Fig. 1), is shown in Fig. 7. The delay PDF results from timestamps obtained by tracing at the application interface and the rising edge of the transmission power at the air interface. Thereby the packet size varies from 1 to 1500 Bytes. The delay distribution is concentrated between 5 to 15 ms, where it exhibits uniform character. This can be explained as a contribution of 5 ms caused by the USB-modem due to data processing and a random contribution of up to 10 ms while waiting for a transmission window. The start of a transmission can only take place at the beginning of a TTI, whereas data appears randomly in the transmission queue and hence, is kept for a random time until the outset of the next TTI. This delay contribution can be removed by designing HSUPA aware software applications. The small amount of packets yielding a delay below 5 ms are measurement artifacts. They result from retransmitted packets or control information, misinterpreted as part of the user data. Increasing the packet inter-arrival time would improve the situation but, as explained in Section 2.1, this would increase the probability of switching to normal Dedicated Channel (DCH) operation. Figure 11(a) shows a histogram of delay and size for $\Delta1$. In contrast to Fig. 7, the falling edge of the transmission power is used to obtain the timestamps, thus, transmission delay is included as well. We model the delay contributed by the UE, the queuing and the transmission as

$$\Delta1 \;=\; 5\text{ms} + X \cdot 10\text{ms} + \lceil l/\alpha \rceil \cdot 10\text{ms} \;, \tag{1}$$

whereas X is a uniformly distributed random variable between 0 and 1, l is the payload length of the transmitted packet, α denotes the length-factor, equivalent to the step hight in Fig. 11(b) (e.g. 800 Bytes) and $\lceil\cdot\rceil$ is the ceiling operation.

NodeB. In Fig. 8 the reader finds a logarithmic histogram of delay $\Delta2$ and packet size, thus, corresponding to the delay introduced by the NodeB. Thereby the value assigned to different colors of the grid corresponds to the natural logarithm of the number of packets corresponding to one parcel of the grid. The delays are calculated by subtracting the timestamp of the falling edge of the

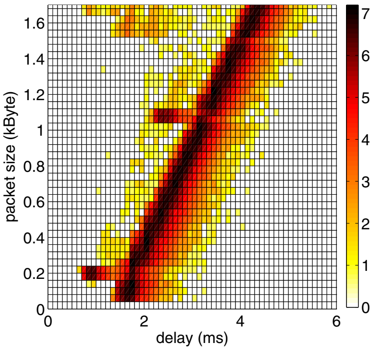


Fig. 9. Delay $\varDelta 3$ over packet size, log. histogram

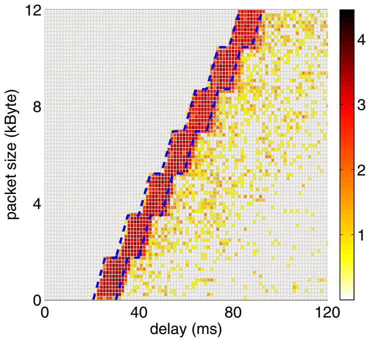


Fig. 10. $\varDelta 1 + \varDelta 2 + \varDelta 3$ over packet size, log. histogram

transmission power at the air interface from the last RLC frame transmitted over the Iub interface. The minimum is 5 ms, whereas up to 7 ms of latency are experienced depending on the packet size. The contribution of the NodeB can thus be modeled as

$$\varDelta 2 \;=\; 5\text{ms} + (l\%\alpha) \cdot 1\text{ms/kByte}\;, \tag{2}$$

whereas % denotes the modulo operator and the expression $(l\%\alpha)$ introduces extra delay growing linearly with packet size.

RNC. Figure 9 shows the delay characteristics of the RNC, $\varDelta 3$ (see Fig. 1). The delay is the difference in time of the last RLC packet fragment at the Iub interface and the last IP packet fragment at the Gn interface.

The minimum latency introduced by the RNC is 1.5 ms. Additionally the packets experience an extra delay up to 4 ms, depending on the packet size. This can be modeled in the same way as for the NodeB by

$$\varDelta 3 \;=\; 1.5\text{ms} + (l\%\alpha) \cdot 2\text{ms/kByte}\;. \tag{3}$$

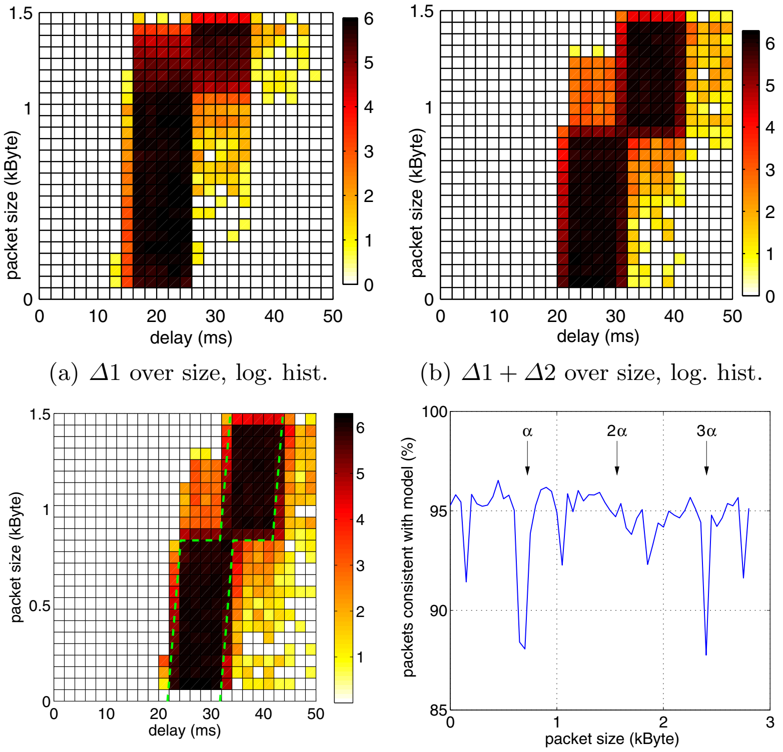
3.2 Accumulated Delay

Accumulated delay is the delay experienced by the user application. Figure 10 displays this accumulated delay for a large variation of packet sizes. The dashed lines correspond to the model illustrated below. Furthermore, Fig. 11 shows the accumulation of the latency throughout the 3G network. In other words those figures show the delay contributed by the first *hop* of the communication link.

By combining the Eqn. (1), (2) and (3), we obtain an expression for the accumulated latency,

$$\varDelta \;=\; 11.5\text{ms} + X \cdot 10\text{ms} + \lceil l/\alpha \rceil \cdot 10\text{ms} + (l\%\alpha) \cdot 3\text{ms/kByte}\;. \tag{4}$$

This expression accurately models the regions of high density in Fig. 10 and 11(c). The parameter α is strongly dependent on a manifold of parameters, such as data



(a) $\Delta 1$ over size, log. hist.

(b) $\Delta 1 + \Delta 2$ over size, log. hist.

(c) $\Delta 1 + \Delta 2 + \Delta 3$ over size, log. hist.

(d) Model evaluation

Fig. 11. Accumulated delay

rate and channel quality, and is defined by the HSUPA scheduler. Comparing Fig. 10 and 11(c) we estimate α as 1800 Byte and 800 Byte respectively, which is a considerable variation, although in the measurement setup only the mean data rate changed. All possible values for α are listed in [21], Annex B. Nevertheless, we expect α not to drop below 200 Bytes for reasonable channel conditions.

The accuracy of the model can be visually evaluated from Fig. 10 and 11(c). The dashed lines show the lower and upper bounds, within which the model assumes a uniform distribution. In Fig. 11(d) a numerical evaluation of the accuracy is depicted. Thereby a data set of packet sizes up to 3 kByte is compared to the model. The result shows that 90 - 95 % of all packets are consistent with the model.

4 Conclusion and Outlook

In this paper we present measurements, analysis and models of latency components of 3G HSUPA communication. We inspected the network elements - user equipment (UE), base station (NodeB) and network controller (RNC) - in live operation. The average delay value for a 1kByte packet is 30 ms (UE: 66%, NodeB:

20%, RNC: 14%). Therefore the 3GPP Long Term Evolution (LTE) delay performance target of 5ms makes improvements in the core network mandatory. Based on the analysis of the results we designed a model for each of the delay components. It provides an average performance of 95%.

Acknowledgments. We would like to thank Sebastian Caban, Robert Langwieser and Michael Fischer for contributing technical expertise. Furthermore, we thank *A1 Telekom Austria* and *ftw* for technical assistance and the *EU FP7 LoLa Project* for financial support.

References

1. Holma, H., Toskala, A.: Hsdpa/Hsupa For Umts. In: High Speed Radio Access for Mobile Communications. Wiley, Chichester (2006)
2. LoLa consortium. D2.1 Target Application Scenarios (2010), `http://www.ict-lola.eu/`
3. Arlos, P., Fiedler, M.: Influence of the Packet Size on the One-Way Delay in 3G Networks. In: Krishnamurthy, A., Plattner, B. (eds.) PAM 2010. LNCS, vol. 6032, pp. 61–70. Springer, Heidelberg (2010)
4. Fabini, J., Karner, W., Wallentin, L., Baumgartner, T.: The Illusion of Being Deterministic – Application-Level Considerations on Delay in 3G HSPA Networks. In: Fratta, L., Schulzrinne, H., Takahashi, Y., Spaniol, O. (eds.) NETWORKING 2009. LNCS, vol. 5550, pp. 301–312. Springer, Heidelberg (2009)
5. Fabini, J., Wallentin, L., Reichl, P.: The importance of being really random: methodological aspects of IP-layer 2G and 3G network delay assessment. In: ICC 2009, Dresden, Germany (2009)
6. Paxson, V., Almes, G., Mahdavi, J., Mathis, M.: Framework for IP Performance Metrics (1998), `http://www.ietf.org/rfc/rfc2330.txt`
7. Liu, J., Tapia, P., Kwok, P., Karimli, Y.: Performance and Capacity of HSUPA in Lab Environment. In: VTC Spring 2008, Singapore (2008)
8. Vacirca, F., Ricciato, F., Pilz, R.: Large-Scale RTT Measurements from an Operational UMTS/GPRS Network. In: WICON 2005, Budapest, Hungary (2005)
9. Romirer-Maierhofer, P., Ricciato, F., Coluccia, A.: Explorative analysis of one-way delays in a mobile 3G network. In: LANMAN 2008, Cluj-Napoca, Romania (2008)
10. Option Wireless Technology, `http://www.option.com/`
11. LoLa consortium. D 3.2. Network related analysis of M2M and online-gaming traffic in HSPA (2010), `http://www.ict-lola.eu/`
12. Endace DAG, `http://www.endace.com/`
13. The Darwin Project, `http://userver.ftw.at/~ricciato/darwin/`
14. Ricciato, F.: Traffic monitoring and analysis for the optimization of a 3G network. IEEE Wireless Communications 13, 42–49 (2006)
15. 3GPP. TS 25.401, UTRAN overall description, `http://www.3gpp.org/`
16. Linear Technology, LT5534 - RF Power Detector, `http://www.linear.com/`
17. SiRF star III GPS Receivers, `http://www.csr.com/products/technology/gps`
18. libpcap - library for network traffic capture, `http://www.tcpdump.org/`
19. LinuxPPS Project, `http://wiki.enneenne.com/index.php/LinuxPPS_support`
20. Vienna University of Technology, Institute of Telecommunication - Downloads, `https://www.nt.tuwien.ac.at/downloads/featured-downloads`
21. 3GPP. TS 25.321, MAC protocol specification, `http://www.3gpp.org/`