

ARQ Concept for the UMTS Long-Term Evolution

Michael Meyer, Henning Wiemann
Ericsson Research
Ericsson GmbH
Aachen, Germany

Mats Sågfors, Johan Torsner
Ericsson Research
Oy LM Ericsson AB
Jorvas, Finland

Jung-Fu (Thomas) Cheng
Ericsson Research
Ericsson Inc.
Rayleigh, NC, USA

Abstract— Work is ongoing in 3GPP to significantly extend the performance of UMTS in the work item Long Term Evolution (LTE). LTE contains a new system architecture with fewer infrastructure nodes and it has been decided to terminate the ARQ functionality in the evolved Node B. This paper describes the requirements that exists for an LTE ARQ concept and outlines a solution that fulfills these requirements in the new LTE architecture. The solution builds on two layered ARQ feedback mechanisms that provide sufficient reliability with a low resource cost for the feedback. The paper contains thorough stepwise argumentation how we arrived at the proposed ARQ concept.

Keywords- *Wireless Link Layer, ARQ, HARQ, LTE*

I. INTRODUCTION & BACKGROUND

Currently, the downlink performance of most WCDMA-based 3G networks is upgraded to support high-speed packet access (HSPA) which includes high-speed downlink packet access (HSDPA) defined in WCDMA Rel 5 and high-speed uplink packet access (HSUPA) defined in WCDMA Rel 6. Both provide significant performance enhancements to 3G networks, in terms of peak data rates, latency and capacity. The Long-Term Evolution (LTE) is the next step in the UMTS evolution roadmap. The targets set by 3GPP [1] include peak data rates higher than 100 Mbps in the downlink and 50 Mbps in the uplink (in a 20 MHz spectrum), further reduced latency and further increased capacity compared to WCDMA Rel 6.

It has been agreed in 3GPP that the LTE/SAE system architecture will differ from the current UMTS network architecture. Only two infrastructure nodes remain: the evolved NodeB (eNodeB) and the so-called access gateway (AGW). However, since essentially the same functionality as in previous UMTS releases needs to be supported by the LTE concept, a re-grouping of functionalities that formerly resided in radio network controller (RNC), serving GPRS support node (SGSN) and gateway GRPS support node (GGSN) has to take place. This paper discusses the consequences this architectural modification has on the automatic repeat request (ARQ) concept and closely related link layer functions like segmentation and data multiplexing. A more general LTE system overview can be found in [2].

The ARQ functionality in HSPA is distributed in two protocol layers (radio link control (RLC) and medium access control (MAC)) as well as across three nodes (user equipment (UE), NodeB and RNC). The hybrid ARQ (HARQ) protocol is part of the MAC layer and is terminated in the NodeB and UE.

On the other hand, ARQ is contained in the RLC protocol that is terminated in the UE and RNC. The HARQ protocol is responsible for handling transmission errors by performing retransmissions based on HARQ schemes such as incremental redundancy or Chase combining. The RLC ARQ has two tasks. First, it is responsible for correcting residual HARQ errors, i.e., transmission attempts where the HARQ process could not deliver the data successfully. Second, in WCDMA the RLC ARQ supports lossless inter-NodeB cell changes, since it is terminated in the RNC that is not affected by the cell changes.

One of the architecture decisions made in 3GPP was to exclude ARQ functionality from the AGW and to terminate ARQ in UE and eNodeB. This architectural decision leads to the question whether LTE should use only a single (H)ARQ protocol instead of a two-layer approach with HARQ and ARQ on top. Furthermore, in LTE there is no ARQ protocol that can be used for lossless hand-over support since the protocols are terminated in the eNodeB. Consequently, other means for providing reliability in hand-over situations are required.

In the present paper, we show that achieving the required reliability with a single HARQ layer can be very costly in terms of resources needed for HARQ feedback. Thus, we argue that a two-layered ARQ/HARQ approach is the best way of achieving both high reliability and low resource cost for the ARQ/HARQ feedback. We also elaborate on specific mechanisms for protocol interaction between ARQ and HARQ facilitated by the fact that both layers are terminated in the same nodes.

The paper is outlined as follows: In the next section, the main requirements are introduced that need to be considered for the ARQ concept design. Section III discusses the two alternatives of using only a HARQ layer or two-layer ARQ design. Based on the conclusion that it is preferable to use a tightly coupled two-layer ARQ approach, the characteristics of this design choice are described in detail. Section IV explains the interactions between HARQ and ARQ and Section V presents the proposed cell-change procedure.

II. REQUIREMENTS

This section discusses different requirements that should be considered for the design of the LTE link layer concept. First we discuss several general requirements. In a second step, we quantify the reliability requirement for TCP.

A. General Requirements

Reliability: The main purpose of ARQ protocols is to deal with transmission errors so that protocols higher up in the stack perceive a basically loss-less channel. The required level of reliability depends on higher layer protocols or applications and will be discussed in more detail below.

Delay: For a well-performing system it is important that delays are minimized. In the ARQ protocol context this applies especially to how quickly retransmissions can be done and to the use of timers.

Resource Cost: It is vital that the ARQ concept is cost-efficient in terms of resources needed for both the data transmission itself and the feedback path. High reliability and low delays should be provided without sacrificing extensive resources. It is also important that the ARQ concept operates resource efficient across the whole range of data rates.

Simplicity: Complexity should only be increased if justified by significant gains, since complexity leads to increased cost in development, testing, and operation. In particular, performance tuning often leads to significantly increased complexity and thus needs to be treated with care.

Scalability: LTE aims at supporting a wide range of applications. While VoIP applications require only a few kbps, high-speed file downloads may use data rates of more than 100 Mbps.

In-Order Delivery: In particular, TCP has quite strict requirements for the in-order delivery of higher layer data units. Although, work is on-going to make TCP less sensitive, it is still required to re-order IP packets that are shuffled due to ARQ retransmissions.

B. TCP's reliability requirement

It is well-known that TCP puts stringent requirements in terms of delays, in-order delivery and packet losses [3]. At high bit-rates, it is fair to state that TCP is one of the most demanding *applications* when it comes to the required packet loss rate. Therefore, the LTE ARQ concept should be designed in order to meet those requirements. Fig. 1 shows the achievable TCP throughput as a function of the packet loss rate for two different fixed network delays (one-way) of 10 and 25 ms between eNodeB and file server. The end-to-end delay can be determined as the sum of this fixed delay and the transmission delay over the air interface including the impact of required HARQ retransmissions. For the air interface, a HARQ error rate for the first transmission of 20% is assumed, while the Transmission Time Interval (TTI) is set to 0.5 ms and the HARQ Round Trip Time (RTT) to 3 ms. The results are obtained by an analytical tool similar to the one described in [4], which comprises models for the HARQ protocol and TCP. The size of the downloaded file is 10 Mbytes.

Fig. 1 depicts the TCP throughput for a radio bearer rate of 100 Mbps, which corresponds to the LTE requirements in [1]. The packet loss rate should clearly not exceed 10^{-5} for the optimistic case of 10 ms fixed network delay, while loss rates below 10^{-6} are required for 25 ms fixed network delay. Otherwise, the packet losses perceived by TCP and the corresponding TCP congestion control actions will prohibit

full utilization of the available bearer resource. This means the requirement for the LTE network set-up on residual errors is stringent.

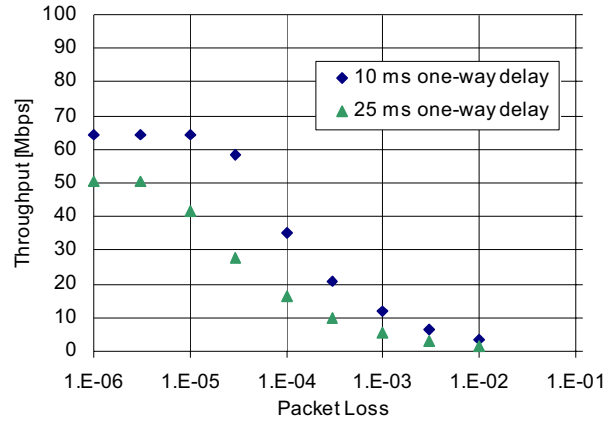


Figure 1. TCP throughput as a function of the packet loss rate for two fixed network one-way delays and a radio bearer rate of 100 Mbps. The difference to 100 Mbps at low loss rates is due to TCP slow-start, HARQ retransmissions and protocol overhead.

However, TCP does not require that the packet loss rate is zero, since TCP provides its own ARQ mechanism and below 10^{-6} packet loss rate no significant impact on the performance can be expected.

III. PROTOCOL DESIGN

A. HARQ and HARQ feedback

In cellular systems, it is state-of-the-art to use a HARQ protocol to achieve a high system capacity and the use of HARQ is also an undisputed assumption for LTE. HARQ typically supports fast feedback to provide low delays for retransmissions, and to reduce the size of HARQ buffers for soft-combining.

Considering that (H)ARQ in LTE is terminated in the eNodeB, it looks appealing to use only a HARQ protocol to ensure reliability. A second protocol may add complexity, additional control signaling overhead, and in the worst case harmful protocol interactions. However, this approach is not as straightforward as it seems due to several reasons.

By allowing a large enough number of retransmission attempts the HARQ protocol can be configured to achieve the required residual packet loss rate (e.g. 10^{-6}). However, the robustness of the HARQ feedback mechanism is critical to obtain this error rate. Since the end-to-end RTT needs to be low to achieve a high throughput, it is desirable to have fast and frequent HARQ feedback to correct transmission errors as soon as possible. It is natural to consider the same approach used in HSPA, where a synchronous one-bit ACK/NACK signal is sent every transmission attempt and the timing of the feedback message is used to identify the corresponding data transmission. That solution achieves the fastest possible feedback while minimizing the information in the feedback message. However, this binary feedback is susceptible to transmission errors and, in particular, NACK reception errors (i.e., erroneously interpreted as ACKs at the receiver) lead to data losses at the HARQ layer. Thus, the reliability of the

HARQ layer is bounded by the error rate of the feedback and not the error rate of the data transmission. However, it is costly to achieve a sufficient HARQ feedback reliability for the frequently transmitted feedback message. The transmitter power requirements are high because of channel fading and the constraint that a single bit cannot be protected with good Forward Error Correction (FEC) codes. Instead, one has to rely on repetition and energy accumulation within the delay limit of the synchronous protocol to combat fading. It is therefore imperative to consider alternative solutions to trade off resource cost, coverage and ARQ feedback reliability.

One solution for achieving a higher reliability without excessive expenditure on HARQ feedback is to apply a second layer of ARQ on top of the MAC HARQ layer, e.g., to use an RLC acknowledged mode as is done for HSPA. Note that in this set-up the second ARQ layer is mainly responsible for correcting error events due to HARQ feedback errors and not transmission errors. One main benefit of the additional ARQ protocol is that it provides a much more reliable feedback mechanism based on asynchronous status reports with explicit sequence numbers that are protected by a cyclic redundancy check (CRC). This implies that the receiver of the status report can detect any errors in the report through the CRC. The reliable transmission of the feedback information is further enhanced in several ways. First, the status messages are protected by turbo codes. Secondly, HARQ is also applied to status messages. Thirdly, the status messages are accumulative. Even if the transmission of a status report fails, the subsequent status includes the information of the last one.

B. Performance Evaluation

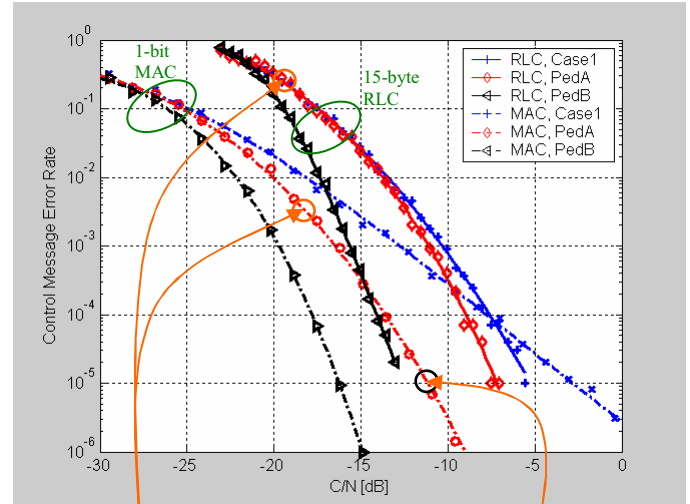
In the following we evaluate two protocol concepts and compare them in terms of their required C/N to meet the desired reliability level of 10^{-5} .

Alternative 1: Only a HARQ Protocol is used: A one-bit HARQ feedback is sent after every detected transmission attempt. The required error rate for a NACK-to-ACK error is 10^{-5} . It is assumed that the initial error rate for the first HARQ data transmission is 25%. Thus, in 75% of the cases an ACK is sent, while only in 25% the more problematic case, a NACK, occurs. In this example, to achieve a NACK-to-ACK error rate of 10^{-5} an error rate of 4×10^{-5} is thus required.

Alternative 2: An ARQ protocol is used on top of the HARQ protocol to deal with residual HARQ errors, e.g. due to NACK-to-ACK errors. Again, HARQ feedback is sent after each detected transmission. In this case, it is assumed that the HARQ feedback has a reliability of only 10^{-3} . For the simulation results presented below, an RLC status message size of 15 bytes was assumed.

Fig. 2 shows the carrier to noise ratio (C/N) required to achieve certain error rates for the two alternatives. The figure contains plots for both alternatives under for three different fading channel models: 3GPP case 1, Pedestrian A and Pedestrian B channel profiles. The vehicular speed is set to 10 km/hr. Each MAC HARQ feedback bit is repeated eight times and then sparsely placed across the entire spectrum to gain frequency diversity. The 15-byte RLC message is protected by a rate 1/5 turbo code. The results for both RLC and MAC

message error rates include losses due to realistic channel estimation. In the following, we shall use the mildly dispersive Pedestrian A channel results (red curves) as the primary numerical example. For this channel, a C/N of -12.5 dB is required for alternative 1 to reach the desired feedback error rate of 4×10^{-5} (or 10^{-5} NACK-to-ACK error rate). For alternative 2, a C/N of -18 dB is sufficient to reach an error rate of 4×10^{-3} for MAC HARQ and a C/N of -19 dB is needed to reach a BLER of 25% for the first HARQ transmission of RLC status reports. The more demanding C/N for alternative 2 is the one for HARQ feedback: -18 dB, i.e., a difference of 5.5 dB compared to alternative 1.



Two-layered ARQ
 $RLC_{err} = 2.5 \times 10^{-1} \rightarrow -19$ dB
 $MAC_{err} = 4 \times 10^{-3} \rightarrow -18$ dB

MAC HARQ alone
 $MAC_{err} = 10^{-5} \rightarrow -12.5$ dB

Figure 2. Numerical results of the needed C/N to achieve various message error rates for 1-bit MAC feedback messages and 15-byte RLC status reports

The present example clearly illustrates that the HARQ feedback is very costly compared to an in-band ARQ feedback mechanism based on cumulative, asynchronous status reports. To achieve high reliability, low delays and low cost, we therefore propose to combine the two.

Internal simulation studies have shown that the different flavors of the LTE link layer ARQ concept that are currently being discussed at 3GPP show very little difference in performance. Differences are more subject to chosen simulation scenarios than on protocol functionality itself. The simplification of protocols to reduce costs for development, testing and proper configuration appears to be more important than providing tweaks and optimizations for rare events. However, of course the main performance requirements need to be met. The simplicity versus performance trade-off requires careful considerations and complexity shall only be increased if it is clearly justified by significant gains.

C. Protocol Design

Based on the findings presented above, we can conclude that it is beneficial to operate an ARQ protocol on top of the HARQ protocol, because this approach requires a lower C/N and thus saves transmission resources and is beneficial to

increase the cell coverage. In fact, it would be sufficient to use just a second reliable feedback mechanism within the HARQ protocol to solve the problem discussed above, but there are more arguments why the back-up ARQ mechanism is operated as a separate protocol layer. These are discussed in the remainder of this section.

The first reason to operate the ARQ protocol in a separate protocol layer is the requirement to multiplex different radio bearers into one transport block. These radio bearers have typically different QoS requirements. For example, the system could be configured such that VoIP packets are not subject to ARQ retransmissions if the HARQ process fails to deliver the IP packet. On the other hand, TCP segments require a higher reliability and are retransmitted by ARQ. With a two layer ARQ concept, each flow can have a separate ARQ entity allowing that ARQ retransmissions could be done independently. However, data from both flows would be part of the transport block that is subject to the HARQ transmissions. In other words, the HARQ retransmission unit is the transport block with data from potentially several radio bearers while an ARQ Protocol Data Unit (PDU) is the ARQ retransmission unit.

Another reason to separate the ARQ layers is in-order delivery. HARQ shuffles the order of PDUs since a varying number of retransmissions might be required for some of the transport blocks. Many protocols or applications can not cope well with out-of-order packets, e.g., TCP. Thus ARQ protocols have to provide in-order delivery. To allow for in-order delivery per radio bearer, a sequence number is required per bearer. The most straightforward solution is to use the ARQ sequence number for this.

The considerations provided in this section lead to the following ARQ concept proposal for LTE:

- HARQ handles transmission errors and uses binary, synchronous feedback.
- The HARQ retransmission unit is a transport block that may contain data from more than one radio bearer (MAC multiplexing).
- ARQ handles residual HARQ errors, i.e., it retransmits data for which the HARQ process failed.
- ARQ retransmission unit is an RLC PDU.
- RLC performs segmentation or concatenation according to scheduler decisions. An RLC PDU contains either a segment of a Service Data Unit (SDU), a complete SDU, or may contain data of several SDUs (concatenation).
- In case of no MAC multiplexing, there is a one-to-one mapping between HARQ and ARQ retransmission unit.
- RLC performs in-order delivery to higher layers.

IV. HARQ-ARQ INTERWORKING

In the previous section, the overall ARQ concept has been derived and the responsibilities of the HARQ and ARQ layer have been laid out. However, the exact interworking between the two requires further consideration. In HSPA, the HARQ

and ARQ protocol operate basically independent of each other. The different protocol termination points on the network side (ARQ in RNC, HARQ in NodeB) do not provision for a tight coupling between the two. For LTE, the situation is different. Both, HARQ and ARQ are terminated in the same nodes, in UE and eNodeB. This allows for a tighter interworking.

What are the advantages of a tighter interworking? First, ARQ protocols often use timers to protect against certain deadlock situations. In WCDMA, the configuration of these timers is a complex task since the network architecture might be very different from one network to another resulting in different delays. Thus, the timers often need to be set conservatively to cope with worst case scenarios. This may slow down the protocol operations unnecessarily. If protocols are operated in the same node, it is more efficient to use triggers to inform an upper protocol of certain events rather than relying on upper layer timers. This eliminates delays that are due to conservative timers. Second, and this is the more important reason, even if two protocols are specified, the implementation of these two protocols might make use of certain shortcuts. For example, the same memory could be used to store data units. Also protocol states can be shared easily.

What are the situations, where a tight Interworking is beneficial? In more than 99% of the cases HARQ deals with transmission errors without the need to engage ARQ. Only in the case of residual HARQ errors, ARQ has to react. There are three such error cases.

1. Maximum number of HARQ retransmissions is reached and the HARQ process is terminated.
2. NACK-to-ACK error as described above. Data is lost.
3. DTX-to-ACK error: A transmission was scheduled, but the receiver did not detect the associated control information and is therefore not aware that anything was intended for it. It does not send HARQ feedback, but still the sender believes that it received an ACK. Data is lost.

For all three cases, an independent ARQ protocol has to rely on its own mechanisms to detect such errors. For example, in WCDMA, RLC has to detect a missing PDU or it has to wait for the expiry of a poll timer before it can react. Both, lead to unnecessary delays for the error recovery.

For LTE, we therefore propose to use triggers from the HARQ protocol to the ARQ protocol as soon as HARQ sender or receiver is aware of the residual error.

For the first error case above this is straightforward. The HARQ sender knows that the transmission failed and that no further retransmission is allowed. Thus, it triggers an ARQ retransmission.

The second error case is detected by the HARQ receiver, because it expects another HARQ retransmission but it receives a new transmission. In this case the receiver sends two feedback messages. The first is the ordinary HARQ feedback for the new transmission. In addition, it sends an explicit ARQ feedback message indicating the residual HARQ error. This procedure assumes that the HARQ sender is not allowed to suspend an on-going transmission. Note that the HARQ receiver does not necessarily know the RLC sequence

number(s) of the data contained in the missing transport block that was subject to the NACK-to-ACK error. When a new transport block is sent, the HARQ-related control information sent out-of-band identifies that a new transmission occurred. The receiver then knows that a residual HARQ error occurred, but it does not know which RLC PDUs were affected. In order to minimize retransmission delays, the receiver should report in the reliable feedback message an explicit timing reference of the missing transport block instead of waiting until the RLC sequence number is known, e.g., when the current transport block is successfully decoded. To report the timing reference is therefore a faster error recovery mechanism. It is also simpler to use the timing reference instead of checking all on-going ARQ and their states for missing PDUs. On the other hand, at the sender it is simple to map a timing reference to the PDUs that were contained in the respective transport block. These can then be retransmitted by ARQ. Overall, this is an example which highlights the benefits of close HARQ-ARQ interworking.

The third error case (i.e., when the receiver failed to detect the transmission) can not be detected at the HARQ layer, because the receiving HARQ entity is not aware that it missed anything. However, the ARQ receiver can detect this event as a missing PDU in its sequence number space. In contrast to the NACK-to-ACK error case, here the ARQ sequence number has to be used in the reliable feedback, because the receiver has no timing reference to the failed transmission. This means that in the proposed concept the reliable feedback should use two different addressing schemes to point to residual HARQ errors: explicit timing references in case of NACK-to-ACK errors and explicit sequence numbers in case of DTX-to-ACK errors.

The discussion of the HARQ error cases is slightly simplified due to space limitations. A more thorough discussion can be found in [5].

In summary, the first two residual HARQ errors can effectively be corrected by HARQ triggers to the ARQ layer. The third event needs to be handled by ARQ itself. For all three cases it is assumed that the HARQ state for the failed HARQ process is flushed. That means the ARQ retransmission appears as a new transmission for the HARQ. In particular it might be possible to perform a re-segmentation for the ARQ PDU, e.g., if the radio conditions changed dramatically.

V. INTER ENODEB CELL CHANGE

Since the ARQ protocol is terminated in the eNodeB, it can not be used for lossless mobility support and an additional mechanism is required for this. Two alternatives have been debated in 3GPP. In both scenarios, it is assumed that the HARQ state is flushed at cell changes. The first option is an ARQ context transfer, i.e., the complete ARQ state including the buffers is transferred from one base station to the other. This allows that the ARQ can basically continue in the new cell. Such an approach is complex and therefore, 3GPP selected a second, simpler solution: In the downlink, all not acknowledged ARQ SDUs are forwarded to the new base station via an interface between the base-stations denoted as X2. In the uplink, all not acknowledged SDUs are

retransmitted towards the new eNodeB. In order to avoid wasting radio resources by performing unnecessary retransmissions, the MAC scheduler tries to complete all on-going HARQ processes before the connection in the old cell is released. In some cases SDU reordering occurs during the handover. That requires that the original order is re-established. In the downlink, this is done preferably in the new eNodeB, which receives forwarded packets from the old cell but also new data from the AGW. A sequence number used for ciphering and assigned in the AGW can be re-used for that purpose. In the uplink, the AGW can re-use in a similar fashion the ciphering sequence number assigned by the UE. In this way a handover fulfilling the reliability requirements can be performed.

VI. CONCLUSIONS

The first details for the Long-Term Evolution of UMTS are currently being standardized. This paper discussed the relevant design issues for the LTE link layer ARQ protocols. The LTE architecture foresees that the radio link ARQ is terminated in the eNodeB and the UE.

High-speed file downloads require at least packet-loss rates in the order of 10^{-5} to 10^{-6} . Binary HARQ feedback (ACK/NACK) is either not robust enough at a fair level of transmission power or too resource-consuming at this required feedback reliability target. Thus, the detailed analysis leads to a not so obvious conclusion: It is advantageous to employ a two-layered ARQ concept provided that harmful cross-layer interactions are eliminated by tight interworking between the two protocols. This conclusion holds also when the two protocol layers are terminated in the same nodes. A HARQ protocol with fast feedback should be used to handle the main share of transmission errors. In addition, another ARQ protocol is responsible for correcting those few residual HARQ errors to meet the packet loss requirements for high speed TCP-based applications. Only relying on a HARQ protocol without reliable feedback mechanism is not radio-resource efficient.

Furthermore, the paper provided detailed proposals for HARQ and ARQ interworking. The analysis revealed that most residual HARQ errors are detected at the HARQ layer. In addition to the binary HARQ feedback, the HARQ layer should therefore also be able to send reliable feedback messages. Alternatively, these reliable feedback messages can be triggered by the HARQ and transmitted at the ARQ layer facilitated by tight protocol interworking.

REFERENCES

- [1] 3GPP: "Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN)", TR25.913, Rev 7.3.0, March 2006.
- [2] H. Ekström, et al.: "Technical solutions for the 3G long-term evolution," IEEE Commun. Mag., March 2006, pp. 38-45.
- [3] H. Inamura et al., "TCP over 2.5G and 3G wireless networks," IETF RFC 3481, Feb. 2003.
- [4] J. Peisa and M. Meyer, "Analytical model for TCP file transfer over UMTS," in proc. 3G Wireless 2001, San Francisco, CA, USA, May 30-June 2, 2001, pp. 42-47.
- [5] 3GPP R2-061398, "HARQ-ARQ Interactions", RAN WG2 #53, Shanghai, China, 8.-12. May 2006.