DIMENSIONING STUDY FOR ROAD USER CHARGING

Charlott Eliasson and Markus Fiedler
Blekinge Institute of Technology
School of Engineering
Dept. of Telecommunication Systems
371 79 Karlskrona, Sweden
{charlott.eliasson, markus.fiedler}@bth.se

ABSTRACT

Road User Charging (RUC) solutions based on the concept of thin clients rely upon timely delivery of Heavy Goods Vehicles (HGV) positions towards a RUC provider, which then performs the map matching and the calculation of the road user charges. In order to avoid data loss and late delivery, a proper dimensioning of the system is required. For the Swedish case, we investigate the feasibility of legacy mobile and wireless systems (in particular GPRS, UMTS and WLAN) for both streaming and bulk transfer of positioning data, and we determine sensible combinations of sampling and reporting intervals in terms of efficient and economic reporting. We also highlight a couple of traps regarding the dimensioning, in particular of sending patterns that might cause overload at the server side of the system. It becomes obvious that mobile networks can well cope with the extra load caused by reporting HGVs, while bulk transfer in wireless hot-spots can be used as a complementary solution.

Keywords: Road User Charging, Dimensioning, GPRS, UMTS, WLAN

INTRODUCTION

The Swedish ARENA project aims at the design of a Swedish Road User Charging (RUC) system (1). In particular, the Swedish system shall cover the whole national road network and allow for a diversification of road charges amongst other depending on the particular road an HGV uses at a particular date. This implies the need for discovering the positions of a HGV and matching them towards maps and pricing schemes. Such matching can take place on the On-Board Unit (OBU) in case map and pricing information are available; in this case, we speak about a *thick client*. An alternative solution, the *thin client*, relies upon delivery of positions towards a central server operated by a RUC provider that performs the map matching and the calculation of the road user charges. More background information and discussion of these approaches are found in (2) and (3), respectively. Interested readers are referred to the final report of this study (4).

The thin client solution is dependent on timely delivery of positions from the OBU to the server. Loss of this positioning data should be avoided, as it constitutes the basis for road user charges. As up to one hundred thousand HGVs might provide their positioning data from their OBU towards the provider in more or less regular intervals, potentially using the same access network resources as mobile users of voice and mobile Internet service or as wireless users in (public) hotspots, such a solution needs to be robust and cost-effective. Thus, a proper dimensioning of such a system is mandatory.

On this background, this paper will present a study of dimensioning issues for RUC. It will also provide quantitative dimensioning guidelines for a Swedish RUC system, given the use of legacy wireless and mobile networks (in particular GPRS, UMTS and WLAN) for reporting HGV positions towards a centralized server. For this purpose, two main scenarios were considered, streaming and bulk transfer. *Streaming* implies reporting small portions of position data on a regular basis. *Bulk transfer* addresses the transfer of potentially large amounts of accumulated position data. Both scenarios rely on authentication and encryption facilities in order to avoid modification of the sensitive position data. The dimensioning focuses on the capacity requirements for transferring the position data for a given number of active HGVs, which is independent of the access technology under consideration. The magnitude of the latter had to be estimated from traffic counts and the assumption of cell sizes in mobile networks, because such information is classified by the operators. The feasibility of legacy mobile and wireless networks is investigated, and sensible combinations of sampling and reporting intervals in terms of efficient and economic reporting are identified.

The paper is structured as follows. First some important parameters are presented and explained. Then the main contributions, which are the case studies of streaming and bulk transfer, will be discussed and the results will be presented. In the end some conclusions and recommendations are presented based on the results from the case studies, and then a short outlook.

KEY PARAMETERS

As a basis for dimensioning, the following key parameters apply to both streaming and bulk transfer scenario:

- The **position sampling interval** t_S denotes the time interval between two successive readings of the GNSS position. We will consider intervals of 1 s, 4 s, 10 s, 30 s and 1 min.
- The **position reporting interval** t_R denotes the interval between two reports of accumulated positioning data. We will consider intervals of 1 min, 5 min, 1 h, 8 h, 1 day and 3 days.
- The **length of a positioning sample** l_{pos} is given by the data format. We will consider a common length of 10 B for longitude, latitude and time stamp. In order to save precious bytes, these values are proposed to be coded in a differential way with regards to reference values.

From these parameters and observing the worst-case assumption that any HGV samples and reports using fixed time intervals, we derive the number of positioning samples per report $p = t_R / t_S$ and the total length of the positioning report data $P = l_{pos}\, p$.

In the case of bulk transfer, we even have to consider the **time horizon** t_H, which denotes the time a HGV is given to complete its transmission of positioning data. Figure 1 illustrates the relationship between t_S, t_R and t_H. The different colors symbolize data from different reporting intervals that are to be sent in the corresponding next interval. Obviously, the time horizon t_H is reasonably upper-bounded by the reporting interval t_R.
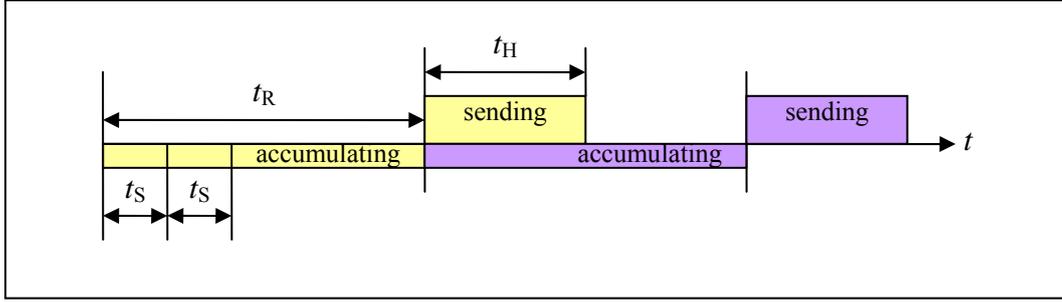
Figure 1. Relationship between sampling interval t_S, reporting interval t_R and time horizon t_H.

STREAMING

In the streaming scenario, we will focus on the required capacity $C(N)$, given in bit per second (bps), kilobit per second (kbps) or Megabit per second (Mbps), as a function of the number N of HGVs in a cell. We round up the observations presented in (4, Section 5.3) to the next order of magnitude, i.e. consider $N_{max} = 100$ HGVs as the maximal number of HGVs per cell. This helps us to easily scale down the results to the number N of interest.

UDP (User Data Protocol) will be used as the transport layer protocol for transferring the positioning data from the HGV via Public Land Mobile Network (PLMN), Gateway GPRS Support Node (GGSN) and Virtual Private Network (VPN) towards the server. The throughput on the VPN is then given by the sum of the throughputs of all reporting HGVs, where the number of 100000 is taken as an upper bound for the fleet of reporting HGVs in Sweden.

We limit ourselves to such combinations of sampling interval t_S and reporting interval t_R that avoid segmentation into several UDP datagrams. If a position reporting was segmented into n datagrams and only one datagram got lost, the reconstruction of this data would need a major selective retransmission and post-processing effort; alternatively, all datagrams belonging to the damaged message are resent. From that point of view, it is much more efficient to have a one-to-one relationship between message and datagram. If a message gets too long, it is better to use a smaller reporting interval instead.

The application data consists out of p positioning samples, for which we assume a length of $l_{pos} = 10$ B, and an identification (ID) number of length 4 B. The calculation of the length L of a positioning message and of the maximal number of samples in order to avoid segmentation is found in Appendix A. We obtain the required capacity $r = L / t_R$ as average over the reporting interval ($\Delta T = t_R$).

Role of the reporting interval

In the case of $t_R = 5$, the sampling intervals $t_S \in \{4\text{ s}, 10\text{ s}, 30\text{ s}, 1\text{ min}\}$ can be applied without the risk of segmentation ($p \leq 60$). In the following Figure 2, results for the sampling interval $t_S = 1$ s are indicated by a dashed line. For $t_S = 4$ s, 100 HGVs produce slightly more than 2 kbps.
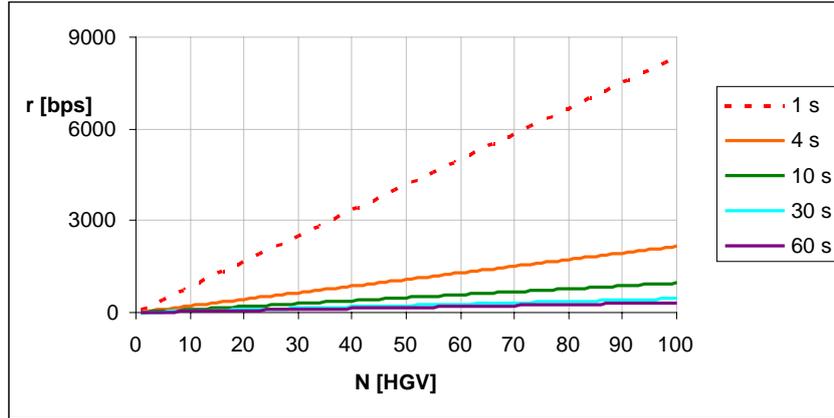
Figure 2. Required capacity versus the number of HGVs for a reporting interval t_R = 5 min and various sampling intervals.

Summary

For the cases avoiding segmentation, Table 1 compares the following parameters:

- *Relative overhead $O = L/P - 1$*, which illustrates the relative part of the overhead ($L - P$) as compared to the amount of positioning data P and thus provides an efficiency index of the reporting as such;
- *Data rate* per HGV r, which provides an idea about the average contribution to the system load;
- *Volume $V = L \cdot 240$ h / t_R* per month when assuming 240 h reporting time per month, i.e. 30 days à 8 h or 20 days à 12 h; and
- *Cost $C = V$ / ($2.70 /MB) per month, assuming a typical Swedish professional user price for mobile data traffic.

Table 1. Different performance parameters of combinations of sampling and reporting intervals avoiding segmentation.

t_S	1 s	4 s		10 s		30 s			1 min		
t_R	1 min	1min	5 min	1 min	5 min	1 min	5 min	1 h	1 min	5 min	1 h
p	60	15	75	6	30	2	10	120	1	5	60
O	11 %	47 %	8 %	107 %	21 %	360 %	72 %	5 %	660 %	148 %	11 %
r	89 bps	29 bps	22 bps	17 bps	10 bps	12 bps	5 bps	3 bps	10 bps	3 bps	1 bps
V	9.2 MB	3.0 MB	2.2 MB	1.7 MB	1.0 MB	1.3 MB	0.5 MB	0.3 MB	1.1 MB	0.4 MB	0.2 MB
C	$25.00	$8.00	$6.00	$4.60	$2.70	$3.50	$1.35	$2.20	$3.00	$1.10	$0.55

As seen from Table 1, longer reporting intervals (implying more positioning samples per report) lead to less overhead, less volume and less cost. Short sampling intervals (implying a high resolution in time and thus also in space) mean comparably large required capacities, volumes and cost. When looking for a high time resolution combined with moderate cost, the combination t_S = 4 s and t_R = 5 min provides a decent compromise with regards to overhead, capacity needs, volume and cost. In case of a low time resolution (t_S = 30 s...1 min), "streaming" data once per hour reaches a high degree of resource use at comparably very low capacity needs, volume and cost.

Capacity considerations

The combination of $t_S = 1$ s and $t_R = 1$ min yields the maximal required capacity. For $N = 100$ HGVs, it basically matches the maximal available capacity of a GPRS carrier when one time slot and Coding Scheme 1 is available. All other combinations need less capacity. However, due to resource sharing effects, the actually available GPRS capacity might be pretty much smaller, say some few kbps, which is basically reached when using $t_S = 4$ s. Several GPRS carriers, time slots or even networks alleviate this problem.

For 100000 HGVs, we immediately can see that no matter of the reporting or sampling interval, the throughput is upper-bounded by 10 Mbps, which can be considered as a rather modest load on a VPN link.

However, these decent capacity values merely apply if the momentary throughput is found pretty close to its average value. Consider the case of $t_S = 4$ s and $t_R = 5$ min. $N = 100$ HGVs generate 2.2 kbps during 5 min (300 s). If all of them would like to get their data through during the same second, the required speed would raise by factor 300 to 660 kbps during that second. Considering the server side (and thus 100000 HGVs), the corresponding peak capacity requirement would amount to 660 Mbps during that particular second, which would have the same effect on a server or a link as a heavy Denial-of-Service attack.

Recommended sending pattern

In order to avoid such traffic peakedness, it might be advisable to make different OBUs send at different times in a controlled *proactive* way. As the OBUs are connected to GNSS, a time reference is available. The placement of the OBUs and their identity numbers within the country and within the reach of certain access points is a pseudo-random process. Assume a minimal reporting interval covers j seconds. We recommend to split this interval into j slots to let OBU i send in slot $(i \bmod j) \in \{0, 1, …, j − 1\}$. This way, we end up with a pseudo-random discrete-uniform-type distribution of time slots in the access networks and even on the link towards the server.

Yet another *reactive* possibility consists in using a so-called Random Backoff Scheme, *cf.* amongst others (5), which aims at resolving overload conditions by making the unlucky station re-sending the data after a randomized (typically exponentially distributed time). The effects on the required capacity will be described in the next section. However, Random Backoff is known to have an inherent risk of instability in case of high offered loads, which is due to the reactive nature of the method.

Bulk transfer

When HGVs cannot stream their data, they need to save the positions to report until there is an occasion for a bulk transmission. Such an occasion might be (a) a necessary break during which a HGV stands still at a parking place with good mobile or maybe even WLAN connectivity; (b) the need to report outstanding positions when leaving the country, which in Sweden means (i) entering a ferry terminal, (ii) crossing a bridge, or (iii) using one of the few land customs facilities to Norway or Finland, which also can be covered by mobile or wireless networks. So the main dimensioning issue for bulk transfer addresses the capacity $C(N)$ required by N HGVs to complete their transmission of positioning data within a certain *time horizon* t_H, which then has to be related to the offerings of mobile and wireless networks.

As outlined in (4, Section 3.2), bulk transfer will use TLS/SSL and TCP. Similarly to the streaming case, the application data consists of $p = t_S / t_R$ positioning samples, for which we assume a length of $l_{pos} = 10$ B, plus the ID à 4 B. The calculation of the length L of the positioning message is discussed in Appendix B. In the following, we consider the following parameters: (a) reporting interval $t_R \in \{1 \text{ min}, 5 \text{ min}, 1 \text{ h}, 8 \text{ h}, 1 \text{ d}, 3 \text{ d}\}$; (b) sampling interval $t_S \in \{1 \text{ s}, 4 \text{ s}, 10 \text{ s}, 30 \text{ s}, 1 \text{ min}\}$.

Time horizon

On the server side, the number of simultaneous connections might be an issue, which is determined by the time horizon t_H. Assuming 100000 HGVs and a uniform distribution of bulk transfers during the day, we obtain the following figures:

- for $t_R = t_H = 1$ min or 5 min: 100000 simultaneous connections (cf. streaming);
- for $t_R = 1$ h and $t_H = 15$ min: 25000 simultaneous connections;
- for $t_R = 8$ h and $t_H = 15$ min: 3125 simultaneous connections;
- for $t_R = 1$ day and $t_H = 15$ min: 1042 simultaneous connections;
- for $t_R = 3$ days and $t_H = 15$ min: 347 simultaneous connections.

Only the last two values seem to be reasonable with regards to server scalability; the third case could be handled when installing a server park. In the sequel, we consider an adaptation of the time horizon such that a rather unproblematic number of simultaneous connections (here 500) is not surpassed:

- for $t_R = 1$ min, this means $t_H = 300$ ms, which is too short;
- for $t_R = 5$ min, this means $t_H = 1.5$ s, which is hardly possible in practice;
- for $t_R = 1$ h, this means $t_H = 18$ s, which is still pretty short;
- for $t_R = 8$ h, this means $t_H = 144$ s $= 2$ min 24 s;
- for $t_R = 1$ day, this means $t_H = 432$ s $= 7$ min 12 s;
- for $t_R = 3$ days this means $t_H = 1296$ s $= 21$ min 36 s.

As the check-in time at a ferry terminal typically spans one hour, the latter values seem to be feasible in any case. On the other hand, a land border passing might imply the need for a shorter time horizon (e.g. 1 min). Taking the time horizon instead of the reporting interval into account, we obtain the required capacity, $r' = L / t_H \geq r$.

Performance and cost

The following table illustrates performance arising from different parameter choices for the bulk transfer scenario.

Table 2. Required capacity r' per HGV for different combinations of reporting intervals/time horizons and sampling intervals; results for non-practicable time horizons (in brackets) suppressed.

t_S \ t_R / t_H	1 min (0.3 s)	5 min (1.5 s)	1 h 18 s	8 h 2 min 24 s	1 day 7 min 12 s	3 days 21 min 36 s
1 s	N/A	N/A	16.7 kbps	16.5 kbps	16.5 kbps	16.5 kbps
4 s	N/A	N/A	4.36 kbps	4.15 kbps	4.13 kbps	4.12 kbps
10 s	N/A	N/A	1.88 kbps	1.68 kbps	1.66 kbps	1.65 kbps
30 s	N/A	N/A	781 bps	578 bps	558 bps	552 bps
1 min	N/A	N/A	518 bps	304 bps	284 bps	278 kps

Table 2 reveals that the required capacity per HGV is mainly a function of the sampling interval, which determines the gross amount of data to be sent during the day. As the time horizon has been adapted to the reporting interval such that, on average, 500 connections exist simultaneously, the size of the reporting interval plays a subordinate role regarding the capacity requirements. Given a certain sampling interval, the differences in required capacities correlate with the different amounts of overhead, cf. also (4, Table 6.2).

Capacity considerations

The total required capacity for N HGVs uploading their data simultaneously within their time horizon is obtained from a multiplication of the values presented in Table 2 by factor N. The corresponding results are shown in Figure 3.
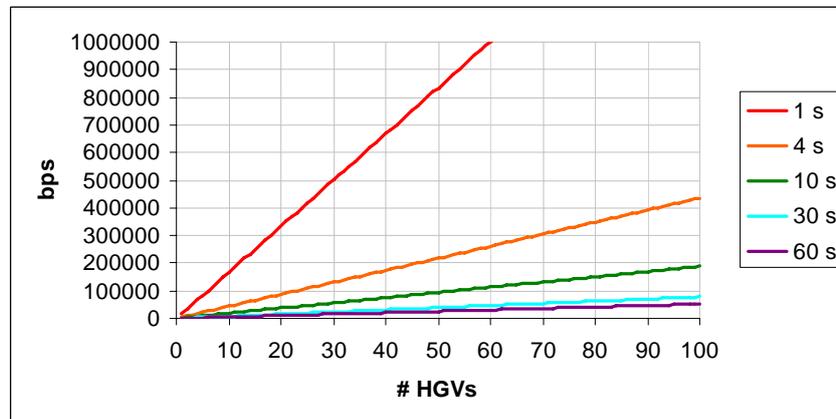

Figure 3. Required capacity versus the number of HGVs for a bulk transfer with a reporting interval $t_R = 1$ h and various sampling intervals.

Given a sampling interval of 1 s, we reach 1 Mbps for 60 HGVs; for a sampling interval of 4 s, we reach almost 450 kbps for 100 HGVs. A WLAN scenario is supposed to be able to handle such capacity demands, in particular because the geographical spread of HGVs implies the need for several access points, offering some tens of Mbps in total.

On the other hand, mobile capacity that typically spans up to several tens of kbps may not be sufficient in these cases involving many HGVs in one spot reporting at the same time. If there

were some few HGVs reporting, they might get own codes assigned in UMTS, yielding guaranteed throughput up to of 64 kbps per HGV. Comparing this to the values reported in (4, Table 6.4), we can see that such a capacity assignment reduces the use of the time horizon by at least factor four. In other words, the HGV uploads and thus releases its connection quicker, which is positive from the viewpoint of server scalability. Some few HGVs could even be handled by GPRS in case large sampling intervals (30 s…1 min) were used.

Recommended sending pattern

Let us turn back to the problem of simultaneous reporting during the time horizon. If 100 HGVs were to report one hour of positioning data each, the time horizon is given by 18 s, see above. Obviously, those 100 HGVs do not need to report at the same time; indeed, their total time horizon amounts to 1800 s, i.e. half an hour, within one hour. That means that if the HGV bulk transfers were perfectly synchronized in time, only one HGV was reporting at a time at maximum. Such a feature would open up for the use of mobile technology (mainly UMTS due to better capacity offerings) even in hotspots. 100000 HGVs and a limitation to say 500 simultaneous connections require 200 time slots within the reporting interval. Following the approach presented for streaming traffic, we let OBU i send in slot (i mod 200) $\in \{0, 1, …, 199\}$. Again, we end up with a pseudo-random discrete-uniform-type distribution of time slots in the access networks and even on the link towards the server, balancing the load in time and space. As outlined above, a potential alternative consists in using a Random Backoff Scheme in order to cope with temporal overload situations.

CONCLUSIONS

From the results presented for streaming and bulk transfer, we can draw a couple of conclusions regarding the dimensioning of the road user charging system:

- Streaming is slightly cheaper than bulk transfer, which is due to less overhead. However, they complement each other; data that cannot be streamed immediately can be stored in the OBU for later bulk transfer at a place with good stationary connectivity (e.g. a parking lot well covered by UMTS, or a ferry terminal offering WLAN connectivity).

- Both streaming and bulk transfer can handle even small position sampling intervals (in the order of seconds). Larger position sampling intervals (in the order of a minute) limit the traffic to such small rates that even GPRS seems to be able to handle the bulk transfer.

- "Heavy" uncoordinated data dumping is only advisable in WLAN hotspots involving several access points. On the other hand, "light" dumping by one or some few HGVs via UMTS and GPRS might work well, given that the coverage is stable.

- In general, the position sampling intervals should be much smaller than the reporting intervals, allowing for a large number of positions per report and thus reducing the capacity-consuming and cost-driving overhead per report. "Good" combinations are for instance a sampling interval of 4 s, combined with a reporting interval of 5 min, or a sampling interval of 30 s, combined with a reporting interval of 1 h.

- The most critical and limiting factor for bulk transfer is the number of parallel sessions at the server, which has to be kept within reasonable limits. Over-dimensioning is no way out, as it still cannot prevent peak loads. Some mechanism controlling the sending pattern or blocking excessive requests will help to avoid overload effects on the server side.

- In order to warrant a continuous flow of data and to avoid excessive resource competition in the access networks and at the server, both streaming and bulk transfers need some kind of time slots assigned to different HGVs, during which the transmissions are supposed to take place. Such time slots could be based upon the identification number of the OBU, leading to a pseudo-random distribution of transmission time slots both across the access networks and towards the server.

<u>Recommendations</u>

From these results, we issue the following recommendations with regards to system design and future work:

- Streaming should be the preferred way of sending positioning data and should be applied whenever possible. It allows for shorter dead times in the payment control loop. Moreover, data that has been streamed successfully does not bother the OBU any more. Accumulated data on the OBU must be seen as some kind of debt within the system. As this debt is growing over time, the need for good connectivity rises in order to "get rid of it" in reasonable time without stressing the system, e.g. by an unnecessarily long connection time towards the server.

- An investigation of the effects of the introduction of time slots for sending data could be studied by means of discrete-event simulation, modeling the data transfers as flows and investigating the load of access networks and server over time.

- Another simulation project might address "heavy data dumping" at a ferry terminal. Here, we could investigate differences between WLAN and UMTS and also investigate the effects of different system parameters (e.g. reporting interval) and control actions (e.g. the introduction of transmission time slots).

- Finally, the implementation of the real system should allow for detailed measurements of network performance and for the collection of end-to-end statistics regarding success and failure of position reporting.

OUTLOOK

Future work should address the validation of the quantitative results and proposals derived in this study. This can be done using simulations ahead of implementation, and using real-world measurements once such a system would be available. Still, the estimation of the number of HGVs, based on statistical properties of HGV traffic, needs to be improved, and corresponding traffic models need to be built. This puts the scene for future research work regarding the dimensioning of a (Swedish) RUC system.

The outcomes of this report are expected to provide aids in determining key parameters and protocols of the Swedish RUC system. As a means of quality control, the implementation of the real system should allow for detailed measurements of network performance and for the collection of end-to-end statistics regarding success and failure of position reporting.

REFERENCES

(1) Homepage of the ARENA project. http://www.arena-ruc.com (last seen 2008-02-13)

(2) J. Sundberg, U. Janusson, and T. Sjöström. A kilometer tax for heavy goods vehicles in Sweden – A conceptual systems design. Part 2: Proposal for systems design. Version 0.93, 2007-01-18.

(3) J. A. Persson, P. Davidsson, M. Boldt, B. Carlsson, and M. Fiedler. Evaluation of road user charging systems: the Swedish case. Proceedings of the 14th World Congress of Intelligent Transportation Systems (ITS 2007), Beijing, China, October 2007.

(4) C. Eliasson and M. Fiedler, Dimensioning Study for Road User Charging, Report for the ARENA project, February 2008. (Please contact authors for digital copy.)

(5) J. Schiller. *Mobile Communications*. 2nd edition. Addison Wesley, 2003.

LIST OF ABBREVIATIONS

GGSN	Gateway GPRS Support Node
GNSS	Global Navigation Satellite System
GPRS	General Packet Radio Service
HGV	Heavy Goods Vehicles
OBU	On-Board Unit
PLMN	Public Land Mobile Network
RUC	Road User Charging
SSL	Secure Sockets Layer
TLS	Transport Layer Security
UDP	User Data Protocol
UMTS	Universal Mobile Telecommunication System
WLAN	Wireless Local Area Network

APPENDIX

A CALCULATION OF L IN THE STREAMING SCENARIO

The length L of a positioning message is obtained as follows: The application data consists out of p positioning samples, for which we assume a length of $l_{pos} = 10$ B, and an identification (ID) number of length 4 B. Authentication is performed by the SHA-224 algorithm [20], which adds 28 B. As encryption, we assume the block cipher Blowfish [21] that needs to apply padding to multiples of 16 B. Finally, UDP overhead amounts to 8 B and IP overhead to 20 B, respectively.

Regarding the limit of the number of positioning samples in order to avoid segmentation, we obtain the following results: Given a typical Maximum Transfer Unit of 1500 B, we arrive at a maximal UDP payload size of 1472 B, which means 92 cipher blocks à 16 B. Substracting 28 B authentication digest and the ID à 4 B, we find room for exactly $p_{max} = 144$ positioning samples à 10 B.

In the absence if segmentation, we can express the length of the whole message, including all overheads, by $L = \text{ceil}(p \cdot l_{pos} + 32 \text{ B})_{16 \text{ B}} + 28 \text{ B} \leq 1500 \text{ B}$, with $l_{pos} = 10$ B, where $\text{ceil}(x)_y$ expresses rounding up the value x to integer multiples of y.

B CALCULATION OF L IN THE BULK TRANSFER SCENARIO

Similarly to the streaming case, the application data consists of $p = t_S / t_R$ positioning samples, for which we assume a length of $l_{pos} = 10$ B, plus the ID à 4 B. The TLS/SSL MAC amounts to 28 B (SHA-224), and a Blowfish encryption implies a padding to multiples of 16 B. The volume of the TLS/SSL handshake is hard to quantify; we assume 400 B on the uplink. Furthermore, TLS/SSL applies a quantization of 16 KB per PDU, i.e. longer messages are split into blocks of 16 KB at maximum. Each TLS/SSL PDU requires 5 + 1 B overhead. On the TCP level, segmentation takes place; given a typical MTU size of 1500 B, we obtain a TCP segment size of 1460 B. Each TCP segment implies 20 B TCP overhead and 20 B IP overhead. In contrast to the streaming scenario, segmentation is not an issue; it is indeed almost unavoidable, given the potentially large amounts of data to be sent. Additionally TCP handshake is assumed to take two packets à 40 B on the uplink. Finally, we arrive at the length L of the positioning data, including all kinds of overheads.