
1 Road pricing as a marketplace

by Werner Brilon

Ruhr-University Bochum, Germany,

e-mail: werner.brilon@rub.de

1. Introduction

All over the world the overload of motorway systems is a problem of high importance. Congestion on motorways is regarded as a significant impediment of economic growth. Due to the increasing division of industrial production processes among many production sites the demand for freight transportation will further increase. Also the demand of personal transportation has a further increasing tendency in growing economies. On the other hand the supply with extended freeway systems is limited due to environmental reasons, lack of space in metropolitan areas as well as due to a lack of funds from public budgets.

The overload of freeway systems is, however, concentrated on specific sections of the network, mainly within metropolitan areas, and on short peak periods during the day. Usually the peak periods cover only a few hours during the whole year. In the most congested part of Germany, e. g. most of the significantly overloaded links of the network suffer congestion below a frequency of 200 h per year. These congestions, however, on a 14 km long section of the Cologne-ring motorway are responsible for 10 Mill. vehicle-hours of delay per year, the economic value of which is estimated as 100 Mill. Euro per year (Zurlinden, 2003)¹.

One of the reasons for the temporary overload of the freeway infrastructure is the fact that every driver is admitted to the infrastructure at any time regardless of the urgency of his trip. On the other hand, within the traffic flow there will be some drivers who would be willing to pay if they could avoid delays due to congestion, whereas other drivers could suffer some delay or could even choose a different time or route for their trips without any harm. In this sense the free availability of highway infrastructure contributes to congestion. This is the case for untolled freeways. But also for tollroads where the toll is charged on the base of kilometers traveled, there is no incentive to avoid peak hours for those drivers who could also use other options (time and/or route) without major problems.

There is no doubt that the use of freeway bottlenecks at a specific time has some kind of economic value. Charging of this value from the road users could be used to avoid congestion on one side and to contribute to funding the infrastructure on the other side. This peak charging of road traffic could either be organized by a fixed structure of peak hour tolls or by a market system.

On the first glance the possibility of a market structure is not visible for a normal freeway. On a closer view we can, however, identify structures which would easily enable to establish a real market mechanism for peak hour charging. These are toll gates near freeway bottlenecks or at the entrance to longer freeway sections. Here the width of the road space usually allows to store the approaching traffic in several lanes where the lanes could be served with different qualities.

As an example one could mention the toll plaza of the Oakland bridge at the east shore of Oakland Bay in California where the approaching traffic is buffered and guided in a maximum of lanes. Each of these lanes could be charged by a different toll. Here the expensive lanes

¹ Tab. 38, S. 163

could be served without any delay whereas drivers on the cheaper lanes could suffer a specific delay due to reduced service rates at the gates. Other toll motorway systems (like in Greece or France) have large toll plazas which would also allow a differentiated tariff for each gate at peak hours.

One other example is the consideration in the Netherlands (Westland, 2000) to construct buffers in the freeway network upstream of significant bottlenecks. This has been studied as a solution to solve congestion problems connected to a tunnel in the freeway network near Rotterdam. Here the idea was to build a huge parking area upstream of the tunnel. Traffic during peak hours was intended to be guided through this area via parallel lanes where each lane has a price, which is indicated at the entry. The high price lanes would serve the vehicles immediately. They are expected to be mainly used by trucks and other commercial vehicles. On the far right side of this space also lanes without any charge could be provided which might, however, induce a significant delay to the vehicles. The gates at the front of each lane are then controlled by algorithms that are based on charged tolls as well as on optimum traffic flow requirements.

Traffic flow characteristics are another aspect which might be served by a peak hour toll charging system. Here it should be observed that the maximum throughput is higher under free flow conditions compared to traffic which is operating under queued conditions. Traffic flow science has clearly pointed out (e. g. Banks, 1991; Hall e.a., 1991; Regler, 2004) that flowing traffic on a freeway provides a higher capacity (= max. flow in veh/h) than the discharge flow at the front of a traffic queue. This difference is called capacity drop. This capacity drop seems to be rather random (Regler, 2004). It could reach an amount of 5 – 15 % of the capacity. Thus, traffic control should try to avoid any breakdown, because congestion induces a reduction of capacity. Statistical considerations reveal that the optimal control would be to keep the traffic volume on a freeway within a range of $\beta = 0,9$ times the capacity (Zurlinden, 2003). This kind of control could also be achieved by an optimized toll gate system. Spiliopoulou e.a. (2008) have shown that tollgates operating on several lanes at a motorway, thus, can contribute to a larger reliability of traffic operations.

Even if the congested periods may cover several hours of the day, the periods where demand is beyond capacity may be very short. If a breakdown due to a short overloaded period once has occurred it might cause a rather long congested period due to the capacity drop effect. Thus, it might become valuable to avoid short overloaded periods at specific bottlenecks by special treatments such as tollgates and buffers in the system.

These considerations give space for a couple of unsolved questions:

- Which would be a practical market mechanism for finding useful peak hour charging?
- Can market price tolls contribute to improved freeway operations? How is this influenced by the capacity of the bottleneck and the demand?
- Can and should these peak charging tolls also contribute to costs of the infrastructure?
- Is this kind of charging for the use of road infrastructure a tool suitable to internalize external costs (like harm on the environment)?
- To which extend can and should the toll income contribute to the revenues of the operator?
- How do these demand responsive tolls influence behavior patterns of the road users?

By this study most of these questions can not be addressed. This study is more concentrated on some of the operational issues which are combined with such a peak pricing system.

2. Model

We assume a bottleneck within a freeway which has a limited capacity c [veh/h]. The demand is time dependent. At first we treat a peak period of duration T [h] with a given demand of q [veh/h], which is assumed to remain constant during period T (cf. Figure 1). q could also be larger than c . After the period of duration T we assume a demand of $q_1 = k \cdot q$, where $q_1 < c$ and $k < 1$.

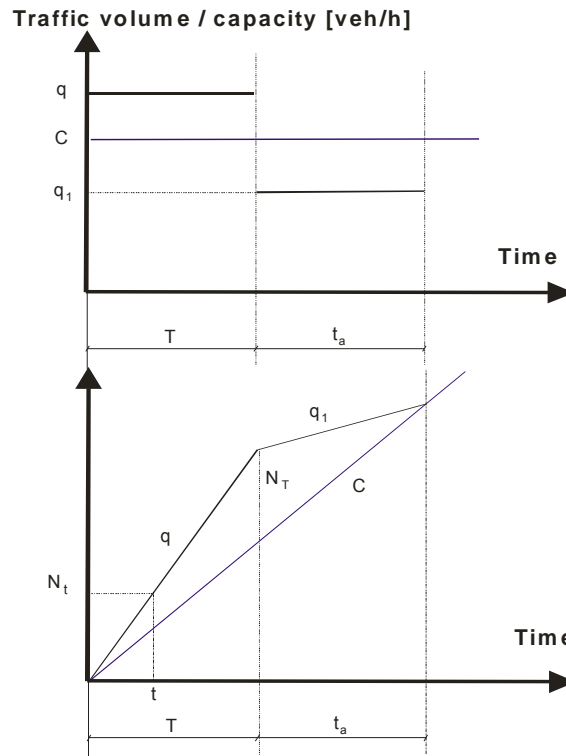


Figure 1: Pattern of traffic demand and capacity over time (upper part) and development of queue length over time

Based on simple deterministic queuing theory the total delay D due to the temporary overload of duration T can be roughly estimated as

$$D = \frac{T^2}{2} \cdot \left[(q - c) + \frac{(q - c)^2}{c - q_1} \right] \quad (1)$$

For the capacity c here the queue discharge capacity should be used which – due to the capacity drop – is lower than the free flow capacity.

The infrastructure installation which we assume is illustrated in figure 2. Upstream of the bottleneck there is a buffer area with parallel lanes, at the front side of which there is a toll gate. This could either be a real gate or also some electronic charging system which uses traffic lights to control the traffic. In any case the possible departure of a vehicle is indicated by a green light and the closure by a red light respectively. At the entrance to each lane we assume that there is an indication to the drivers of the expected delay together with the level of the toll. The lane with the highest toll should be number 1. Overall, we assume that the buffer provides n lanes. We assume that the indication of the expected delay to drivers is

quite reliable. The control algorithm for the green lights should implement features which guaranty this reliability based on queuing theory solutions.

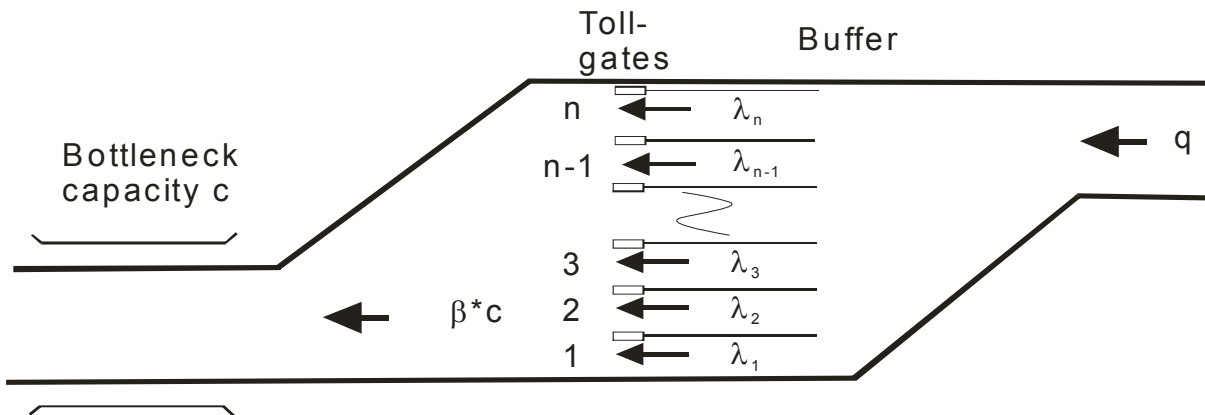


Figure 2: Model of the buffer and toll gates

The delay for a vehicle approaching gate i at any time t is the number $y_i(t)$ of vehicles queuing at gate i divided by the service rate $\gamma_i(t)$ which is applicable at time t .

$$d_i(t) = \frac{y_i(t)}{\gamma_i(t)} \cdot 3600 \quad (2)$$

This is only true for the case that $\gamma_i(t)$ is not changing over time – at least during the clearing of the existing queue. Constant service rates γ_i , thus, are a precondition for the reliable indication of expected delay to the driver. If the service rates would be varied – especially if they would be reduced - while a driver is waiting in the queue then he would experience a larger delay than has been indicated to him at his arrival. Thus, in the interest of the reliability of the system constant service rates for each lane are assumed for this study. However, increased service rates (i.e. reduced delays) may occur as soon as the queues within the system can be reduced due to decreasing demand.

For the first approach we assume only one kind of vehicles, e.g. passenger cars. The vehicles are, however, differentiated according to their drivers' willingness to pay for a trip through the bottleneck with reduced delays.

3. Driver's decision process

The driver when approaching the toll gates gets an indication of the delay and the cost which he has to expect at each of the gates. He has to decide which of the gates he should prefer based on his payoff between costs and delay. For the mechanism of this decision two kinds of models are proposed.

3.1 Logit model

The process of lane selection by the approaching drivers is a choice between discrete alternatives. For this choice a simple Logit model can be assumed. This means: the probability that a driver selects lane i is:

$$p_i = \frac{\exp(V_i)}{\sum_{j=1}^n \exp(V_j)} \quad (3)$$

$$V_i = B_0 + B_d \cdot d_i + B_h \cdot h_i \tag{4}$$

where

B_0, B_d, B_h = parameters of driver behavior

d_i = delay on lane i (s/veh)

h_i = charge for using lane i (\$/veh)

The parameters B_0, B_d, B_h could only be estimated from real world driver behavior in a situation like the one described. As long as this installation is not really available, useful assumptions about these parameters have to be used for sample calculations.

3.2 Cost of delay

Driver's decision can also be modeled as being based on his willingness to pay for reduced delays. The amount of money which the driver is ready to pay for a reduced delay of one hour is called the cost of time. This parameter is assumed to be a random variable which is equally distributed (cf. figure 3) between a minimum W_{\min} and a maximum W_{\max} [\$/h]. Here the minimum could also be zero or even negative for drivers who would prefer a payback in the case that this is offered for a longer delay. The constant distribution is, of course, only a rough preliminary approach. A more realistic distribution might even be rather skewed depending on the proportion of commercial vehicles and the prevailing travel purposes of car drivers. A first approach to the value of time in road traffic might be given by Koenig e.a. (2004). Here the value for conditions in Switzerland varies between 17 and 30 SFr/h.

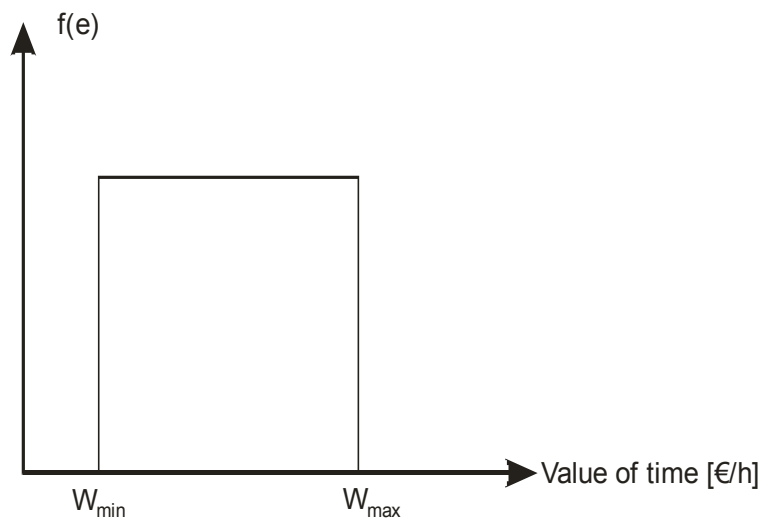


Figure 3: Uniform distribution for the value of travel time.

	Lower margin	Upper margin
commuter	2.9	21.4
shopping	3.7	18.1
business		32.5
leisure	0.8	12.3
All travel purposes	2.1	18.2

Table 1: Value of time: an example from Switzerland given in SFr/h (Koenig e.a., 2004)

4. Solutions of the model

The total demand q , by the toll gate system, is diverted into n streams, each of which has a volume λ_i [veh/h]. Each toll gate should be controlled such that its capacity is γ_i [veh/h].

The different parameters of the model like

lane capacities	γ_i	[veh/h]
demand rates for each lane	$\lambda_i(t)$	[veh/h]
queue lengths for each lane as a function of time	$y_i(t)$	[veh]
delays to individual vehicles or on each lane i at a time t	$d_i(t)$	[s]
charges for the use of lane i	$H_i(t)$	[€]

are related to each other in a rather complex manner. It seems not to be possible to describe these relations by one universal set of equations. Instead, different strategies for the control of the system will be defined. A strategy here means a set of specified and plausible assumptions for relations between some of the variables, which then are constraints for the remaining variables. According to each strategy the development of the remaining variables as a process over time then will be evaluated. Along with this an analysis of sensitivity regarding the different assumptions and parameters should be carried out. Based on an assessment of these results the usefulness of the strategies will be evaluated.

For each of the strategies several basic assumptions are defined:

- Lane n should always be usable without any congestion charges. This seems to be necessary to make such a system to become politically acceptable.
- Lane 1 should always operate without any delay as long as even possible. A capacity $\gamma_1 = 1200$ veh/h as a practical value can be assumed to guarantee an adequate traffic flow quality (i.e. without significant delay; as long as the charging process does not require more than 3 s) on such a lane. The charges on lane 1 should always be determined from the value of time.
- The system might induce a delay to drivers by handling the passage through the system which in the following examples is estimated as 3 s for each vehicle.
- The system always starts with empty queues on all lanes.

5. Strategies

Within these constraints several strategies could be possible. The strategies discussed in the following paragraphs can only serve as examples. More and different approaches seem to be possible.

5.1 Strategy I:

- This strategy uses a decision process of the drivers described by a logit model.
- Also the slowest lane ($i=n$), operating without any charging of the drivers, should maintain a minimum service rate throughout the time, with $\gamma_n = 1/M \cdot \gamma_1$. Thus, also the most slowly served drivers will keep their delays within acceptable margins.
- The service rates γ_i at the intermediate lanes i with index i from 2 to $n-1$ should be determined from a linear interpolation between γ_1 and γ_n such that $\sum_{i=1}^n \gamma_i = \beta \cdot c$. This results to

$$\gamma_2 = \gamma_n + \frac{(c \cdot \beta - \gamma_1 - (n-1) \cdot \gamma_n) \cdot (n-2)}{\sum_{i=2}^n (n-i)} \quad (5)$$

$$\gamma_i = \gamma_n + \frac{\gamma_2 - \gamma_n}{n-2} \cdot (n-i) \quad \text{for } i = 3 \dots (n-1) \quad (5)$$

These service rates are kept constant over time (see exception below).

The delays at any time instant t are computed from eq. 2 for each of the lanes. This delay is offered to each driver who is arriving at the tollgates.

In cases where the capacity γ_i of lane i is not utilized due to the expiration of the queue in lane i then the unused part of the capacity is added to the capacity of lane i+1. This is always performed based on the experience in the previous time slice of 10 s duration. If this happens then the delays on lane i+1 will be reduced compared to the initial expectation and compared to the delay which was indicated to the drivers when they were approaching the queue.

To demonstrate the operation under these assumptions the following example might be helpful:

- n = 10; M = 10
- demand q = 4000 veh/h for T = 10 minutes
- k = 0,5; i.e. a traffic demand volume of 2000 veh/h after the peak
- capacity of the bottleneck: c = 3600 veh/h $\beta = 1$
- Parameters of the Logit-model: $B_0 = 0$ $B_d = -0,01$ $B_h = -0,2$
- (which means that the toll is affecting drives much more than delay)

The following figures illustrate the development of several parameters of the system over time. The total delay over all drivers is resulting to 20,3 h. This compares to a total estimated delay of 17 h which would have happened without the toll system (eq. 1; assumed capacity drop during congestion: 10%) which means that this strategy is not able to reduce the total delay. The sum of all toll charges for this short oversaturation period is 860 \$.

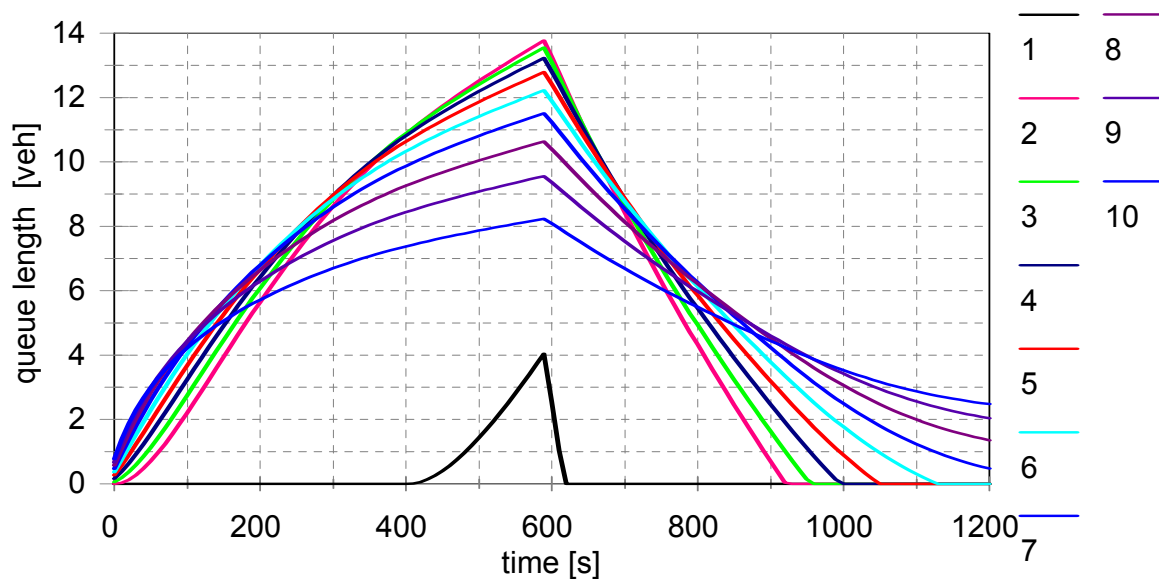


Figure 4: Pattern of queue length over time for each of the 10 service lanes (1 = fast track; 10 = free track)

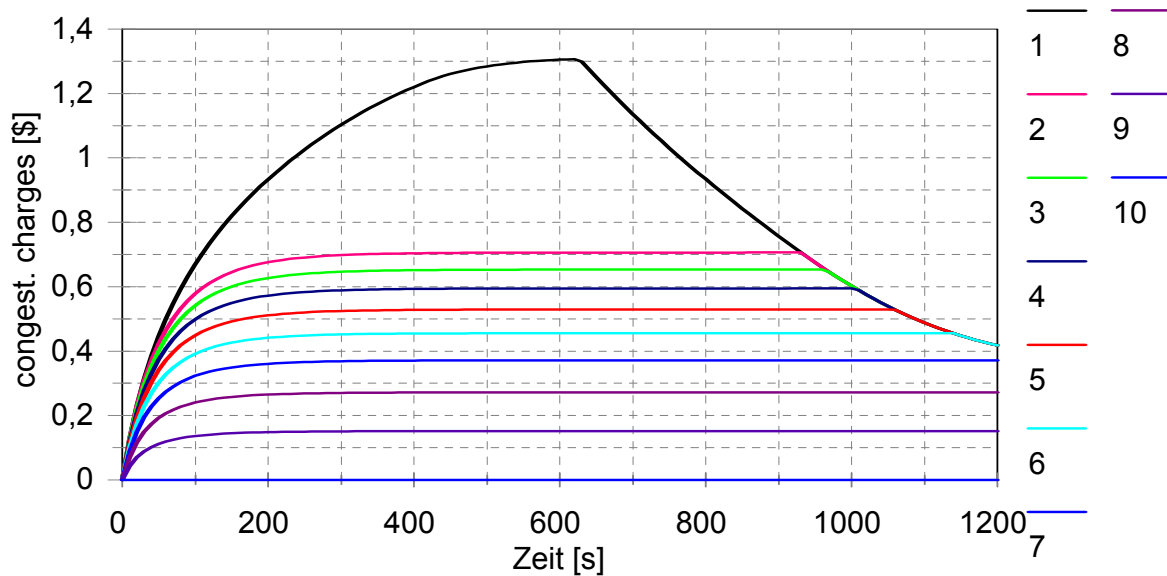


Figure 5: Pattern of the amount of toll per vehicle over time for each of the 10 service lanes (1 = fast track; 10 = free track)

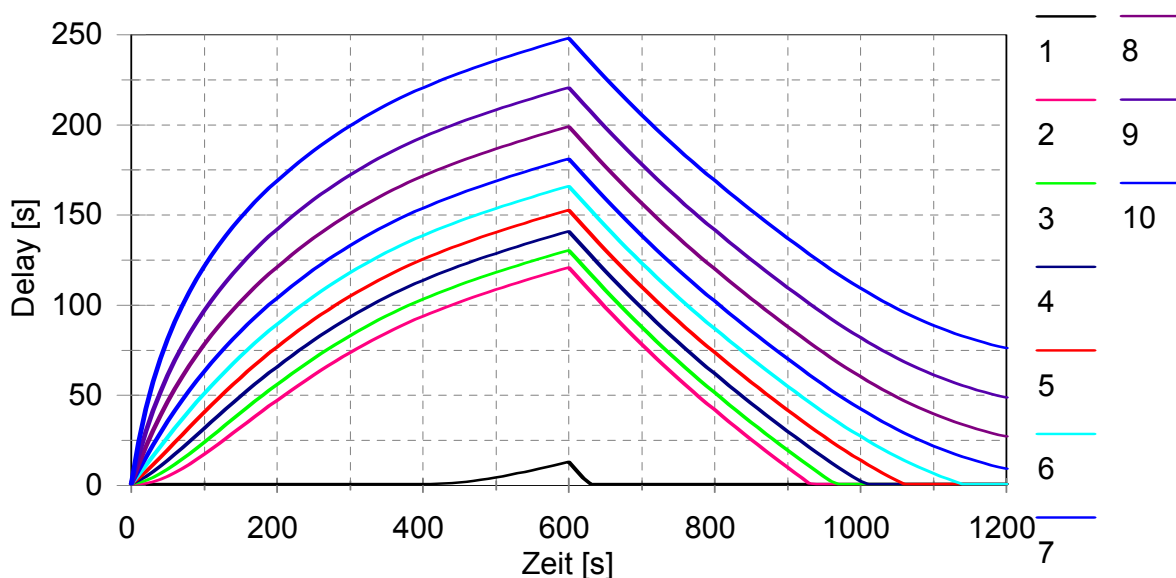


Figure 6: Pattern of the delay to each vehicle over time for each of the 10 service lanes (1 = fast track; 10 = free track)

Figure 5 tells us that the fast lane's toll is growing much higher than at all the other lanes. This is related to the fact that the delay on all the other lanes is significantly larger than on lane 1. On the free lane the delay grows up to 5 minutes at the end of the peak period. Nevertheless, the preferences of the drivers, as expressed by the Logit model, lead to a high acceptance of the fast lane, which is carrying most of the traffic throughout the whole observation period (Figure 7). This does, however, not prevent some drivers to choose the cheaper lanes also from the beginning of the peak period. Thus, they are not immediately served with the consequence that in the initial stage of the peak the existing capacity is not fully used. To avoid this problem the metering system should be started already before the demand reaches capacity. With that precaution the system will be operating on highest capacity when this is needed.

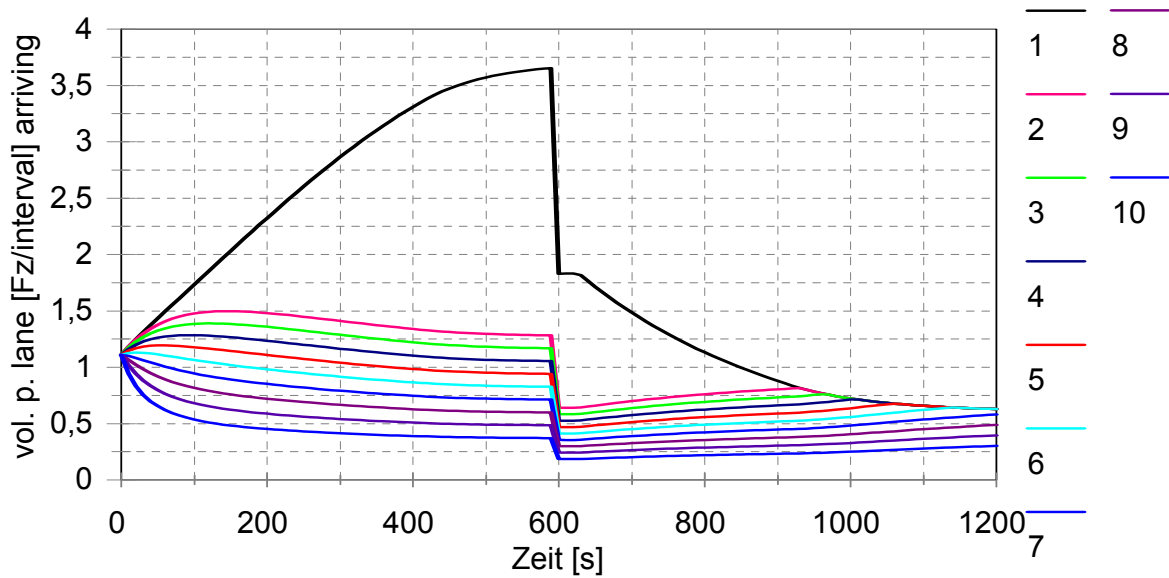


Figure 7: Pattern of the volume over time for each of the 10 service lanes (1 = fast track; 10 = free track)

This would be a reason to modify the strategy. We see also from Figure 4 that the strategy could need some improvements. The fast track 1 is also experiencing some queueing during the final quarter of the peak period. Therefore, the strategy might be modified by offering the same conditions as in lane 1 also for lane 2 temporarily to avoid a queue on the fast track.

Figure 5 tells us that the fees are rapidly growing at the beginning of the metering period. They remain constant after this short time of initiation. This effect happens also with other parameters. Thus, to simplify the system the levels of tolls at the different toll gates should be kept constant from the beginning of the metering period.

5.2 Strategy II:

As one alternative another example for a control strategy is used:

- This strategy uses a decision process of the drivers based on a value of time which is varying among drivers according to a constant distribution in the range of from 0 until 30 \$/h.
- On the fast lane ($i=1$) like in strategy I a capacity of $\gamma_1 = 1200 \text{ veh}/h$ is assumed.
- On the slowest lane ($i=n$) (no toll) – like in strategy I - the capacity is set as $\gamma_n = 1/M \cdot \gamma_1$.
- The service rates γ_i at the intermediate lanes i with index i from 2 to $n-1$ should be following a function of

$$\gamma_i = \gamma_1 \cdot \frac{A}{i} \quad (6)$$

With the restriction of $\sum_{i=1}^n \gamma_i = \beta \cdot c$ we get

$$A_i = \frac{c \cdot \beta}{\gamma_{1,soll} \cdot \sum_{i=1}^n \frac{1}{i}} \tag{7}$$

These service rates γ_i are kept constant over time with the same exception as in strategy I.

The delays at any time instant t are computed from eq. 2. This delay is offered to each driver who is arriving at the tollgates. Drivers are choosing a toll gate according to their value of time.

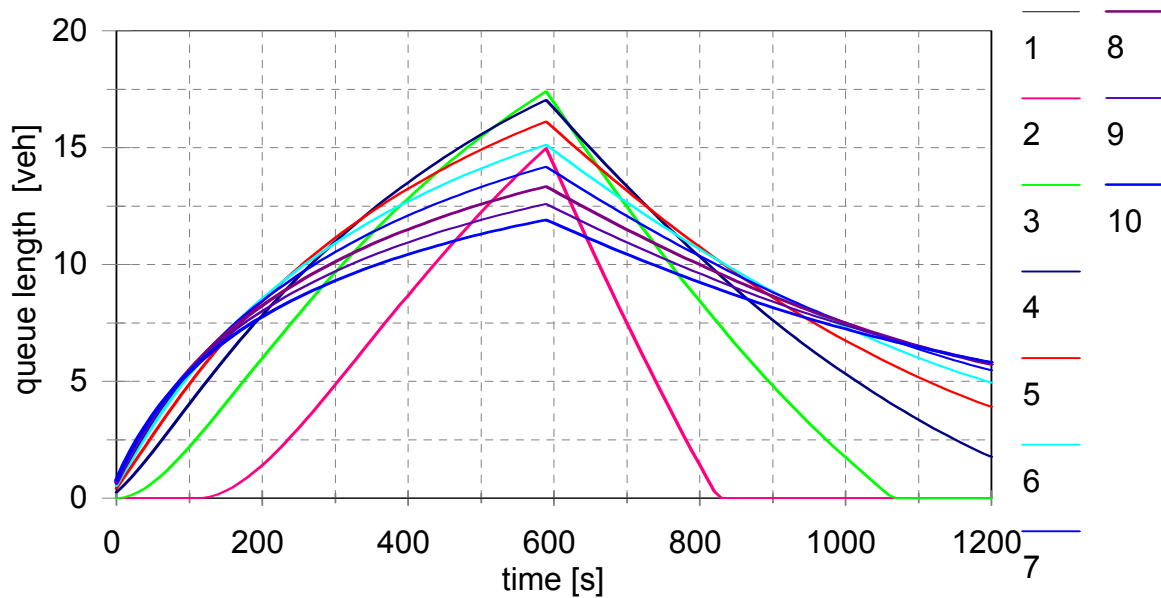


Figure 8: Pattern of queue length over time for each of the 10 service lanes (1 = fast track; 10 = free track)

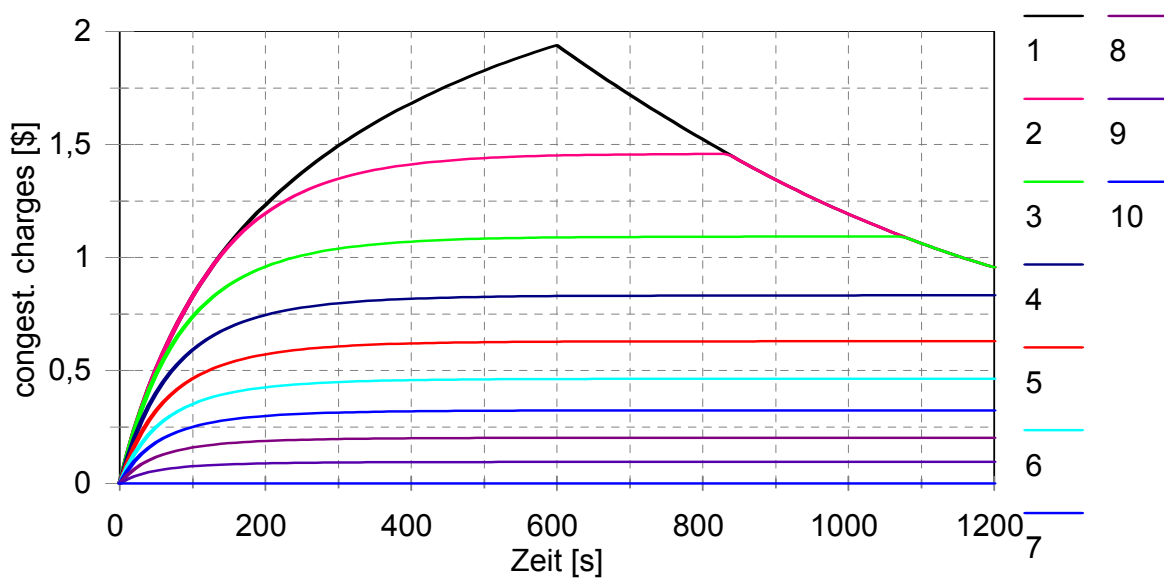


Figure 9: Pattern of the amount of toll per vehicle over time for each of the 10 service lanes (1 = fast track; 10 = free track)

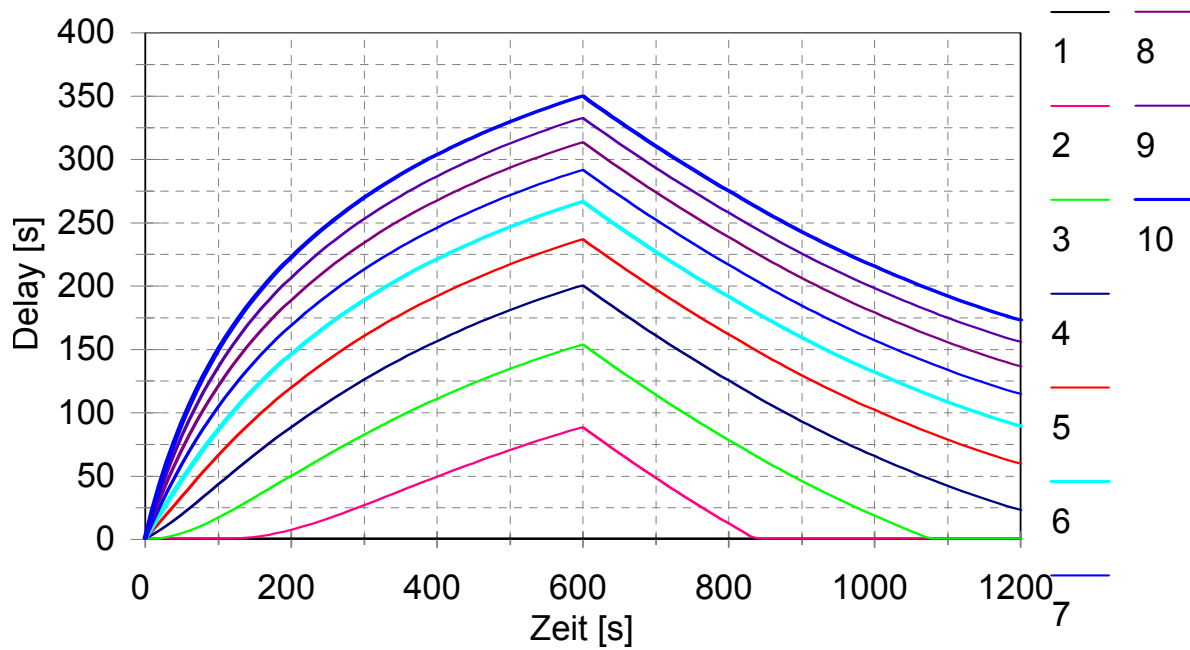


Figure 10: Pattern of the delay to each vehicle over time for each of the 10 service lanes (1 = fast track; 10 = free track)

With this strategy the usage of lanes is more equally distributed regarding the lengths of the queues (Figure 8). Here also delays on the fast track can be avoided. However, the total delay to drivers is 26 h which is more than in strategy 1. Nevertheless the total toll is 905 \$ which is also more than in strategy 1. This shows that the better performance in strategy I is not necessarily combined with a higher toll income.

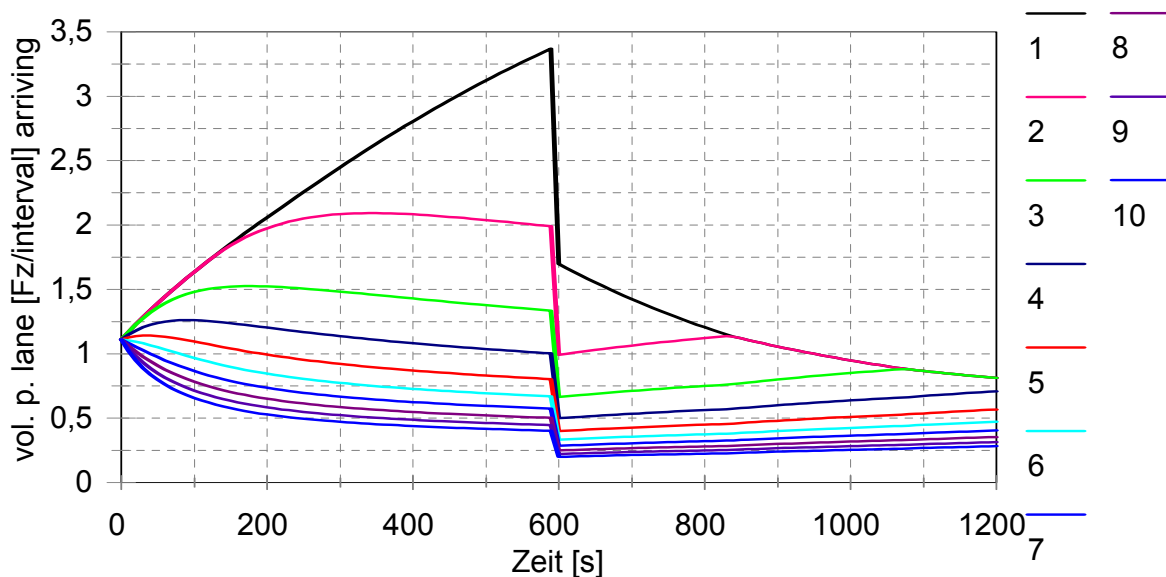


Figure 11: Pattern of the volume over time for each of the 10 service lanes (1 = fast lane; 10 = free lane)

5.3 Other Strategies

One useful strategy, due to reasons of practicability could be to keep queue lengths $y_i(t)$ at any time t equal on all lanes to use the space in the buffer area as efficient as possible. This

could be achieved by keeping the relation of demand for each lane to the reserve capacity of this lane constant over all lanes:

$$\frac{\lambda_i(t)}{c_i(t) - \lambda_i(t)} \approx \text{const. over all lanes } i \quad (8)$$

A control algorithm to make this possible must be adjusted to the driver's acceptance of delays and charges.

As another approach the delay $d_n(t)$ for vehicles using the slowest lane n might be kept in a specific margin compared to the average delay $D(t)$ among all lanes at any time t :

$$d_n(t) \leq \varepsilon \cdot D(t) \quad (9)$$

The intermediate lanes i should provide a delay $d_i(t)$ between $d_i=0$ and $d_n(t)$ according to a specific function

$$d_i(t) = f(i) \quad (10)$$

i.e. for a linear function $f(i)$:

$$d_i(t) = \frac{i}{n} \cdot d_n(t) \quad (11)$$

This strategy still has to be worked out with its consequences.

6. Open questions

This paper should just give a first impression for the operation of a buffer within the freeway system. Of course, there are several unsolved questions. E.g. the technique for collecting the tolls without losing too much time for the drivers still has to be developed. Presumably it should be based on electronic automatic systems. It is also questionable by which kind of system the driver should be given the required information. There is a need for the driver to understand the data instantly and to be able to make a reasonable decision within short time. This makes it also questionable to which degree the information about the delays and tolls should be detailed.

Also the optimization of a control strategy would need a lot of research since there are quite a lot of parameters the consequences of which are highly correlated. An important mechanism is also the way of driver's decision making which must well be modelled to come to the desired optimization results.

7. Conclusions

As a means for the optimized appointment of capacities to drivers at significant bottlenecks within freeway systems road pricing could be used to assign the existing capacity to those drivers for whom the value of additional travel time due to congestion would be the highest. This kind of toll collection could be performed while storing the overflow queue in a buffer. Within this buffer several lanes with different delay to drivers could be formed on which tolls are collected in according to the amount of delays suffered by drivers.

The study tells us that such road pricing systems where the drivers can choose between delays and tolls could become possible. The size of the queues within the buffer would remain limited since an overload occurs only temporarily. The delays to drivers, however, would amount to significant differences to make the system working effectively. The tolls

assigned to each lane should be kept constant over time. It is useful to start the system before demand reaches capacity. Only thus the full capacity will be available at the instant when it is needed. The system has not necessarily the benefit that it reduces the total sum of all delays. This is due to the fact that the system induces some delay to each of the drivers due to the kind of service. Therefore, the main utility of such a system would be that the better performance is appointed to those road users who are willing to pay. This assumes that these are the drivers who can make the most efficient use of the time which they, thus, save.

Such a system would also generate significant amounts of funds which might contribute to finance the infrastructure and the operation of the freeway.

Moreover, beyond the effects studied here, the need to pay tolls during periods of high demand might cause some drivers to shift their trip to another time with lower demand which would be one of the positive effects of such a control system.

8. References

- Banks, J. H. (1990). Flow Processes at a Freeway Bottleneck. Transportation Research Record No. 1287, Transportation Research Board, National Research Council, Washington D.C.
- Hall, F.L. and K. Agyemang-Duah (1991). Freeway Capacity Drop and the Definition of Capacity. Transportation Research Record 1320, Transportation Research Board, National Research Council, Washington D.C.
- HCM (1950): Highway Capacity Manual. Transportation Research Board, Washington D.C., first edition 1950
- HCM (2000): Highway Capacity Manual. Transportation Research Board, Washington D.C., first edition 1950; current edition 2000
- König, A. and K.W. Axhausen (2004) Zeitkostenansätze im Personenverkehr (Value of time in personal traffic), Final report for SVI 2001/534, Schriftenreihe, 1065, Bundesamt für Strassen, UVEK, Bern, Swiss
- Regler, M. (2004). Verkehrsablauf und Kapazität auf Autobahnen (Freeway Traffic Flow and Capacity). Schriftenreihe des Lehrstuhls fuer Verkehrswesen der Ruhr-Universitaet Bochum, No. 28. Bochum.
- Spiliopoulou, A., Papamichail, I. Papageorgiou, M. (2008). Real-time Toll Plaza Management for Throughput Maximization. Paper 08-0771 presented at the TRB Annual Meeting 2008
- Westland, D. (2000). Dimensioning of traffic buffers for regular users changing their demand into a maximum individual delay. Proceedings of the 4th International Symposium on Highway Capacity, Maui, Hawaii, 2000; TRB-Circular
- Zurlinden, H. (2003). Ganzjahresanalyse des Verkehrsflusses auf Strassen (Whole year analysis of traffic flow on highways). Schriftenreihe des Lehrstuhls fuer Verkehrswesen der Ruhr-Universitaet Bochum, No. 28. Bochum.
-