Queue analysis for the toll station of the Öresund fixed link

Pontus Matstoms*

Abstract
A new simulation model for queue and capacity analysis of a toll station is presented. The model and its software implementation (TQA) have been developed for the Öresund fixed link, but the model is general and can be used also for other sites of similar type. Under a given time dependent traffic flow and a set of specified model parameters, the required number of open lanes can be computed and the resulting capacity estimated by discrete event simulation. The paper presents the queue model and shows how it can be used in practice.

1 Introduction

Toll fees paid by the users will finance the fixed Öresund link between Denmark and Sweden, to be opened for traffic in year 2000. On the Swedish side, in Lernacken close to the city of Malmö, there will be a toll station for motorists coming from both directions. Queue and capacity analysis for this toll station and for other toll stations of similar type is the topic of this paper. We have developed a model based on discrete event simulation for queue analysis under assumptions of, for example, the traffic flow and average service time, the number of open lanes, and details in the physical design. In the paper, we briefly describe the model and show how it can be used. For a more detailed discussion, we refer to Matstoms (1999a).

One factor that makes the Öresund case interesting from a capacity point of view is the relatively high toll fee. With lower fees, drivers can simply through some coins in a basket, or pay cash in a manual lane, and drive away often without asking for a receipt. This implies short service times and relatively high capacity per lane. With a more substantial toll fee the use of credit cards and other factors increase the average service time. Also the use of foreign currencies may have a negative effect on the service time.

In the model, we consider two different methods of payment: dynamic payment and traditional payment in automatic or manual lanes. Dynamic payment is primarily for frequent customers. They will have a transponder unit in the car by which valid prepayment can be checked. Special lanes with transceivers communicating with the transponders are used for these vehicles. Other cars, using non-dynamic payment, are directed to a traditional toll plaza with parallel lanes for manual or automatic service. The toll station considered in the model is shown in Figure 1. In the model, it is possible to adjust details in the physical design. The picture given below should just serve as an example and does not follow the detailed design of the Öresund toll station.

* VTI, SE-581 95 Linköping, Sweden. E-mail: pontus.matstoms@vti.se
In the figure, the arriving traffic comes from the right side. There are two entering lanes and exactly how cars, with respect to vehicle type and methods of payment, are distributed on these is determined by road signs. The two lanes on the left and right side of the central toll plaza are for dynamic payment.

![Figure 1](image)

**Figure 1.** The toll station in the presented model. Here, the traffic comes in two lanes from the right side of the picture. The transceiver sections are in pictures marked by short horizontal gray lines.

Let us now consider the left entry lane. The right lane works similarly but is normally used for heavy traffic only. In the left lane there is an exit, around position 300 meters, for cars using dynamic payment. The prepaid fee is then validated in speed (about 50 km/h) around position 200 meters, where the left transceiver section is located. If everything goes normal and the payment is accepted, the driver continues on the highway and returns to normal speed (about 90 km/h). Otherwise, if something fails, he or she is directed to the leftmost non-dynamic lane on the toll plaza.

Drivers using manual or automatic payment end-up on the main toll plaza, with nine lanes for non-dynamic payment. The rightmost lanes might be reserved for traffic from the right entry lane. In the example shown in Figure 1, there are eight lanes for the left entry lane and one for the right entry traffic. Drivers in these lanes are either manually or automatically served; manual lanes by staff and automatic lanes by machines. After the payment, non-dynamic vehicles are directed back to the highway where they are mixed-up with vehicles coming from dynamic payment.

A model like the one presented here is an important tool in the process of planning and designing a toll station, but can also be used for toll stations already opened for traffic. To determine the required number of parallel lanes of a planned station, it is essential to extensively analyze the capacity under different assumptions of the normal traffic flow and very extreme flow levels. By computing the expected queue length and delay under such different assumption, the number of parallel lanes of different types can be settled such that the risk for capacity breakdown stays on an acceptable level. In this process it is also essential to analyze the effect of very extreme flows under short periods and, for example, what happens when failing vehicles temporarily blocks one or more lanes.
Different parts of the toll station are under normal conditions mutually independent and should not interfere with each other. Under very high flows there is, however, a risk for interference that must be avoided. Following Figure 2, this may for example happen if all opened non-dynamic lanes on the toll plaza are full and the length of the common queue before the entry to the toll plaza increases and the entry to dynamic payment gets blocked. Only three of the nine non-dynamic lanes are in this example open and vehicles must wait in a queue to enter the toll area.

![Figure 2. Example of extreme situation where the common queue to the central toll plaza blocks the entry to dynamic payment in the lower left lane. Drivers using dynamic payment are in this case hindered by the non-dynamic queue. (Source: Animation in TQA)](image)

For an existing toll station, the required number of open lanes during different periods of the day, week and year can be determined from expected flow levels. By measuring current traffic flows on the highway prior the toll station and dynamically using the queue model, it is furthermore possible to tune the service level such that lanes are opened and closed in time. Another important application is to use the model for analyzing the effect of additional lanes or new technology. What happens, for example, if the machines used for automatic payment are replaced with faster machines or if the mean service time by other means can be reduced with a few seconds? What impact does that have on the mean delay in queue or the proportion of driver that must queue? Finally, what would the effect of an additional manual or automatic lane be?

Although there is a well-developed theory for analytic queue theory, we use micro simulation based on discrete events to estimate the discussed capacity measures. It is, of course, natural to ask whether simulation or analytical queue theory should be used. In the latter case, it would be possible to compute the expected delay and queue length without uncertainty. By simulation we get estimates with statistical uncertainty. The reason for using simulation is the flexibility in assumptions. It is then possible to make very special assumptions and to analyze different phenomena without noticeably increasing the computational complexity.
2 The Queue model

The queue model is based on discrete event simulation and consists of small submodels and build-in assumptions, for example on details in driver behavior. In this section, we describe parts of this model.

2.1 Basic assumptions

We consider three vehicle classes: cars, lorries/buses and trucks with trailer. Each type with its own parameters (length, speed, acceleration etc) and proportions of drivers using different pay methods. In the simulation, vehicles are generated such that the time between successive arrivals follows an assumed distribution; typically exponential distribution. The total traffic flow, considered as a continuous function $Q(t)$ of time, is approximated by discrete input observations \( \{ Q(t_i) \}_{i=1} \) connected by linear segments. From this curve the mean flow and the average time $D_t$ between arrivals can be computed at any time during the considered time period. When a new vehicle is generated, it is assigned randomly generated basic properties, like vehicle type and method of payment, following proportions given in the model assumptions.

On entry to the toll plaza, drivers using non-dynamic payment choose a currently open lane of the right type; manual or automatic. Following the model, a driver who wants to pay manually always takes a manual lane while drivers accepting automatic payment can choose between the manual and automatic lanes. A reasonable assumption is that drivers take a lane with least number of vehicles in queue. However, in practice lanes tend to be more or less popular compared with other open lanes. Fixed weights have therefore been introduced for each lane and used in the following way: A driver entering the toll station chooses an open lane of the right type such that the product of the queue length (in vehicles) and the lane specific weight is minimized. This means that lanes with low weights can be chosen even if the queue is shorter in another lane. When a lane has been chosen, the driver is assumed to keep it; even if an adjacent lane turns out to be faster.

2.2 Service time

It is assumed that the basic service time follows a certain distribution. Let us, for example, assume that the service time is normally distributed with a given mean $\mu$ and standard deviation $\sigma$. The service times obtained in this way are preliminary in the sense that they can be modified later. First, if a driver has been waiting in queue for more than $t_{\text{prepare}}$ second, then it is assumed that he or she is prepared when reaching the service point. A reduction of $t_{\text{reduction}}$ seconds is therefore made on the preliminary time. Second, if the resulting service time falls below a certain minimal service time, it is truncated to this value.
Assuming normally distributed preliminary service times, it is clear that most drivers get a service time in the interval $\mu \pm 2\sigma$. Service times longer than $\mu + 3\sigma$ will be extremely rare and much longer time will in practice never occur. However, in reality there are certain drivers with extremely long service time; much longer than the above normal distribution in practice can generate. A failing credit card may, for example, give a much longer service time. In the model, we handle this variation in service times by dividing the drivers in two categories; normal and extremes. Normal drivers are handled as described above but for the extreme drivers we use a completely different distribution for service times with much higher mean and standard deviation. The proportion of drivers of the two types is given as model parameters and can be different for different combinations of vehicle class and method of payment.

In reality, it is reasonable to assume a bit longer service times during low traffic periods. Then drivers do not feel a pressure from behind, and every moment may take a bit longer time. In manually served lanes, drivers may also talk to the cashier for a moment. As the flow increases, the service becomes more efficient and the resulting service times are in average reduced. However, at some point, the average service time starts to increase. This may, for example, be due to delays in the administrative transaction systems. In the model, it is possible to include assumptions of the above type. One example is shown in the Figure 3.

![Figure 3](image.png)

**Figure 3.** Example of flow dependent service time.

This effect may, of course, have some influence on the capacity, but the main motivation for introducing it is for the computation of the required number of open lanes, as described in the next section. Due to shorter average service time, a fixed number of lanes can then be used under a moderate increase in the flow.

Beside the actual time for service, it is also important to model the time it takes for the next vehicle in queue to reach the service point when the previous vehicle has left. In the model, this time is computed from the length and acceleration of the two vehicles. Using the notations introduced in Figure 4, the approach time for the vehicle behind the bus can be expressed as

$$t_{\text{approach}} = \frac{2 \cdot (g + I_{F})}{\min(a_{B}, a_{F})}$$
Here, the numerator includes the distance from the queue head to the service point, and the denominator expresses the acceleration for the vehicle approaching the service point.

Figure 4. Notations for the gap between vehicles in queue and vehicle length and acceleration used in the computation of approach time. Here, the bus is ready to leave service and the small car behind is then drive forward to enter the service point.

It follows that the average approach time depends on some particular vehicle properties and the proportions of the different vehicle types. With reasonable values on the length and acceleration parameters, it turns out that the approach time for the combination car-car is about three seconds, while the approach time for a truck with trailer behind another truck with trailer is more than nine seconds. This variation has an important impact on the computed capacity measures.

2.3 Required number of open non-dynamic lanes

Given a certain time dependent variation of the arriving traffic flow, one important problem is to compute the required number of simultaneously open lanes for non-dynamic payment (planning). We have proposed two different approaches for doing this; a static method based on the current traffic flow and a dynamic approach based on the current queue and capacity usage. We now briefly describe the two methods.

Let \( t_{total} \) be the average service time, including the reaction time and the time it takes to reach the service point from the first position in queue. The maximal acceptable traffic flow in one lane is then given by \( q_{max} = \frac{3600}{t_{total}} \) and the maximal flow to the toll plaza with \( n \) open lanes is

\[
Q_{max} = n \cdot \frac{3600}{t_{total}} \text{ vehicles / hour.}
\]

Moreover, let us define the capacity usage \( \rho \) as the average proportion of time that a service point is in use or waiting for an approaching car. This value can be expressed as the ratio between the current traffic flow \( Q \) and the theoretically highest acceptable flow, \( \rho = \frac{Q}{Q_{max}} \). In practice, the capacity usage must not exceed much more than 95%. If the capacity usage is very close to 100%, there is no extra capacity for handling short periods of
higher flow or incidents. When all capacity is used for the current flow, it might be very
difficult to reduce built-up queues. By putting the above relations together, we get the fol-
lowing expression for the number of lanes to be opened under the total flow $Q$ and maxi-
mal capacity usage $\rho$:

$$n \geq \frac{t_{\text{total}} \cdot Q}{3600 \cdot P_{\text{max}}}.$$

Special care is required when using this basic relation under a time dependent traffic flow
and with lanes of different type and with different service times; see Matstoms (1999a).

An alternative to the static approach is to compute the required number of open lanes dur-
ing the simulation. Then we start either with all lanes closed or a number of open lanes
computed by the static method. Lanes are then opened and closed according to a functional
decision rule:

**Open a new lane if:**

The queue length in some lane exceeds $n$ vehicles or $l$ meter.

**Close a lane if:**

The capacity usage during the last $k$ periods is lower than $P_{\text{min}}$.

Under a reasonable choice of parameters, the dynamic and static method normally gives
very similar results. The static method is easier to use since simulation is not required. A
time dependent flow curve can then be entered and the corresponding recommended num-
ber of open lanes computed without delay. On the other hand, there are certain situations
when the dynamic method must be used. Since the static method only takes the current
flow in account, lanes might be closed even if there is a long queue. Using the dynamic
method, lanes are instead opened and closed with respect to the current queues only.

### 2.4 Simulation

Given a time dependent traffic flow and a corresponding curve for the number of simulta-
neously open lanes, simulation is used to estimate certain measures of the capacity during
the considered time period:

- Queue length (in vehicles and meter)
- Delay in queue
- Proportion of driver that must queue
- Resulting service time

Vehicles are generated with time gaps determined from the prescribed distribution and the
current flow. In the model, vehicles then queue and get service following the above type of
model assumptions. Dynamic as well as non-dynamic payment is handled at the same time.
Each vehicle that arrives to the queue and gets service is an observation and is used in the overall estimate of the capacity measures. By a completed simulation we get:

- Point estimates with statistical uncertainty of the above variables
- Time dependent curves of the capacity measures
- Input to animation

Repeated simulations are used to reduce the statistical uncertainty of the point estimates.

### 2.5 Animation

By animation it is possible to analyze the result of a simulation in detail. Vehicles are shown as small moving dots arriving to the toll station, waiting in queue, getting service and leaving. Furthermore, by the animation it is possible to dynamically open and close lanes, to change the type (manual / automatic) of lane and to modify the set of accepted vehicle types. The effect can then directly be seen in term of shorter or longer queues.

The animation routine takes some input data from the simulation, like the arriving times and service time for individual vehicles. Other variables, like the queue length, are computed during the animation. This is made by time driven simulation, which in some way makes the same job as the event driven simulation in the previous step. The results are the same, but the event simulation is much faster. On the other hand, it is always necessary to make time updates in animation and it also makes it possible to dynamically change properties of different non-dynamic lanes, as described above. In this case, the result will not be the same as in the event driven simulation.

The picture in Figure 2 is, as an example, taken from the animation.

### 2.6 Implementation

The queue model presented here has been implemented as *Toll Queue Analysis* (TQA); a computer program for Windows 9x/NT. It is written in Borland Delphi 3 and includes parameter specification of the model, planning, simulation and animation. The simulation performance on a modern PC is up to 500 vehicles per second. A detailed description of the program can be found in Matstoms (1999b), Matstoms (1999c).
3 Numerical example

Finally, we illustrate how the queue model and the associated software can be used in practical analysis. Under some fixed parameter settings and the flow variation given in Figure 5, the system can be analyzed in terms of the expected queue length, waiting time in queue and proportion of drivers that must queue. Here, the number of opened lanes has been computed by the static planning method presented in the last section. In this case, the simulation indicates only moderate queues (not shown here).

Figure 5. Example of traffic flow (vehicles/hour) on the left axis and model computed number of lanes to be opened on the right axis (staircase curve).

In the example, the flow is rather constant during the first hour, then it is more than doubled before it goes back to the initial level. The static planning method proposes three open lanes from the beginning, then up to five and finally back to three open lanes. As an example, let us ignore the increased flow and constantly use three open lanes. The expected waiting time in queue then stays on a moderate level until the flow starts to increase. Then the waiting time increases; first slowly and then faster, as shown in Figure 6.

Figure 6. Waiting time (seconds) in queue as a function of time. Here, the flow is the same as in the previous figure but the number of open lanes is now fixed to three.
References

