The Multiple Single Server Queueing System

Azra Panjwani
Yang Liu
HuanHuan Qi
University of California, Berkeley
Berkeley, CA

Advisor: Jim Pitman

Summary

Our model determines the optimal number of tollbooths at a toll plaza in terms of that minimizing the time that a car spends in the plaza.

We treat the toll collection process as a network of two exponential queueing systems, the Toll Collection system and the Lane Merge System. The random, memoryless nature of successive car interarrival and service times allows us to conclude that the two are exponentially distributed.

We use properties of single server and multiple server queuing systems to develop our Multiple Single Server Queuing System. We simulate our network in Matlab, analyzing the model's performance in light, medium, and heavy traffic for tollways with 3 to 6 lanes. The optimal number of tollbooths is roughly double the number of lanes.

We also evaluate a single tollbooth vs. multiple tollbooths per lane. The optimal number of booths improves the processing time by 22% in light traffic and 61% in medium traffic. In heavy traffic, one tollbooth per lane results in infinite queues.

Our model produces consistent results for all traffic situations, and its flexibility allows us to apply it to a wide range of toll-plaza systems. However, the minimum time predicted is an average value, hence it does not reflect the maximum time that an individual may spend in the network.

The UMAP Journal 26 (3) (2005) 345–354.

General Definitions

The Network: The point at which the car enters the queue for toll collection to the point at which the car is able to drive off with current traffic speed. It consists of two systems of queues.

Toll-Collection System: The point at which cars arrive at the toll-plaza and form queues to the point at which they exit the booth after toll collection.

Lane Merge System: The point at which cars leave the tollbooth to enter the queue to merge back into the tollway lanes, to the point at which they can drive off with current speed.

Single Server Queueing System: A system with one queue and one server.

Multiple Server System: A system with one queue and multiple servers such that a customer has the freedom to choose any server available.

Arrival rate: The number of cars per minute per lane that arrive to a network or system.

Departure rate: The number of cars per minute per lane that depart from a network or system.

Service or processing: The act of toll collection.

Service rate: The number of cars per minute per booth being served.

Merge rate: The number of cars per minute per lane that merge back into the tollway lanes.

Total time: The time for a car to pass through the network.

Optimal time: The minimum feasible total time.

Idle time: The time interval during which the attendant is not serving anyone.

General Assumptions

- Car arrival times are independent, identically distributed non-negative random variables.
- Cars are served first-come-first-served.
- The service times for individual cars are independent, identically distributed nonnegative random variables with no correlation to the arrival process.
- In the long run, the rate at which cars are served is greater than the rate at which cars enter the network; otherwise, there would be infinite queues.

- There is no limit to the number of cars that can enter the network, because from the point of view of the network, the road length is arbitrarily large.
- Motorists tend to join the shortest queue in vicinity; hence, in the long run, the queue length is about the same at every tollbooth.

Table 1.

Table of variables.

Variable	Descriptions
S	The network of the Toll Collection System and the Lane Merge System
S_1	The Toll Collection System
S_2	The Lane Merge System
λ_1	Average car arrival rate per lane to the S_1 queue
λ_2	Average car arrival rate per lane to the S_2 queue
μ_1	Average service rate per lane in S_1
μ_2	Average merge rate per lane in S_2
W	Total expected time spent by a car in S
W_1	Expected time spent by a car in S_1
W_2	Expected time spent by a car in S_2
ℓ	Average length of a vehicle and the safety distance in front of it
ν	Traffic speed on the road, independent of tollbooth collection
n	The number of lanes in a tollway before the toll plaza
m	Number of tollbooths in a toll plaza
k	Number of lanes in a tollway after the toll plaza

Our Approach

We assume that the cars arrive according to a Poisson process. The arrival of a car at a time t does not affect the probability distribution of what occurred prior to t; hence the system is memoryless [Pitman 1993]. A driver's decision to drive on a road at a particular time is independent from that of any other driver; so the time periods between successive arrivals of vehicles are independent exponential random variables. If the tollbooth attendant is idle, the driver "goes into service"; otherwise, the car joins the queue to be served.

Similarly, the server processes cars with successive service times also being independent exponential random variables. From probability theory, we know that the sum of two exponential random variables with rates λ and μ is another exponential random variable, with rate $\lambda + \mu$.

We apply the theory of exponential queueing systems to develop a model that predicts the value of m that minimizes W.

General Model

A queueing system often consists of "customers" arriving at random times to some facility where they receive service. They depart from the facility at the same rate at which they arrive. The network S consists of two systems, the toll-collection system, S_1, where cars arrive and join the queue and the lane-merge system, S_2, where people join the queue to receive the "service" of merging.

Multiple Single Server Queueing System

We employ queueing theory together with continuous-time Markov chains to build our Multiple Single Server Queueing System, based on the following reasoning.

When cars get to a toll barrier, they determine which queue to join. In theory, they can join the shortest queue. In practice, however, they are unlikely to change too many lanes to join a shorter queue if there are other cars on the road. In most cases, they are limited to entering the queue directly in front of them, or a queue to their immediate left and right. Furthermore, under the assumption that the queue lengths are approximately the same for all the queues, they are most likely to join the queue directly ahead.

The process is similar to a single-server queueing system, but the fact that they have somewhat of a choice in choosing the tollbooth also gives this process properties of a multiple server queueing system. However, multiple-server queueing systems allow for only one queue and the freedom to choose any server that is not occupied. Our system does not fall exactly under either one of the two categories; hence, we coin the name "Multiple Single-Server Queueing System" for the systems in our network, which has the following properties:

- It consists of several parallel single-service queues.
- Each queue has a "super server" that has a processing rate of $\mu_1 \times m/n$.

In **Figure 1**, each colored box represents the probability that a car in a lane uses a tollbooth of that color. The bigger the box, the higher the chance of choosing that tollbooth. Most drivers use the tollbooth right ahead of them, though a few would choose the tollbooth to the left or right (with equal probability). The probability that a car uses a tollbooth that farther away is negligible. As we can see from **Figure 1**, the total areas of all the colored boxes representing the probabilities of going through the tollbooths are eventually the same from lane to lane. This implies that the service rate is the same for every tollbooth. Hence, in the long run, each lane is processed at the rate $\mu_1 \times m/n$.

Similarly, the process of waiting in a queue to merge back into the k lanes of the tollway after paying the toll can also be considered as a Multiple Single-Server Queueing System with processing rate $\mu_2 \times k/m$. To allow for more flexibility in our model, k may or may not equal n.

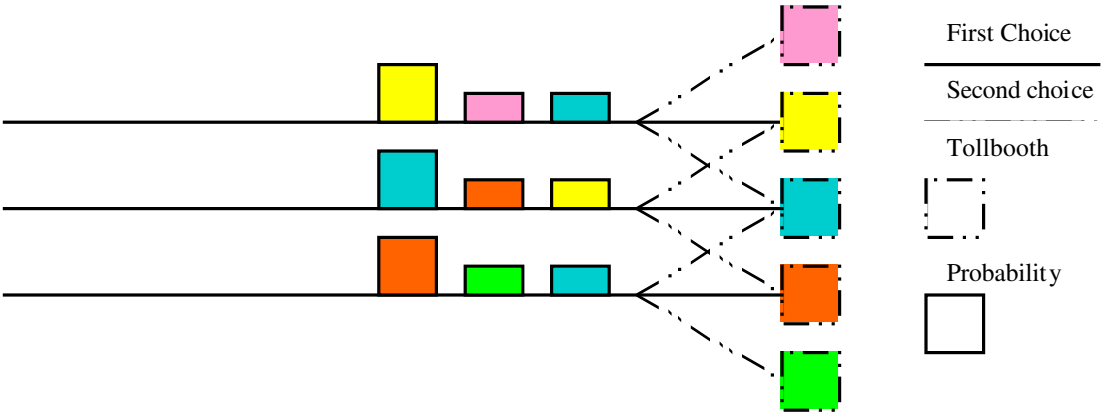

Figure 1. Each lane has equal probability over all cars.

Model Development

The total waiting W is the sum of the times to pass through the two systems, i.e., $W = W_1 + W_2$.

Based on the queueing theory equation [Ross 2003]

$$W = \frac{1}{\mu - \lambda},$$

and the discussion of service rates, we find

$$W = W_1 + W_2 = \frac{1}{\frac{m\mu_1}{n} - \lambda_1} + \frac{1}{\frac{k\mu_2}{m} - \lambda_2}.$$

Derivation of the Service Rates

We assume that on average each tollbooth attendant takes a fixed amount of time t to collect a toll, so $\mu_1 = 1/t$.

For heavy traffic situations, we also take into account driver reaction time r before stepping on the gas and moving up to the booth. We incorporate this delay into the service time to get

$$\mu_1 = \frac{1}{t + r}.$$

We estimate $t \approx 5.5$ s and $r \approx 2.5$ s.

Calculating μ_2 is a little trickier. We take into account ν, which we consider to be determined independently from the toll plaza system. This is justifiable, since whether a toll plaza interrupts a tollway or not, ν varies considerably depending on different traffic situations. Since ν is in miles per hour, and we're interested in cars per minute, we first transform the velocity into meters per minute. We also consider the fact that the car is going from 0 mph to get up to

ν, hence we use the average speed of the car during the time that it must catch up to the tollway traffic. We then divide the velocity by ℓ, which depends on ν, because the safety distance needed for cars at high speed is much greater than that for low speeds. Thus, we obtain

$$\mu_2 = \frac{\nu}{2\ell}.$$

Since cars from the m lanes of the toll plaza must merge back into the k lanes of the highway, we calculate the overall merge rate per lane, $\mu_2 \times k/m$, as described earlier, to be

$$\frac{k}{m}\frac{\nu}{2\ell}.$$

Derivation of the Second Arrival Rate

Since drivers join S_2 as soon as they depart S_1, the rate λ_2 is the same as the departure rate from S_1. Now, consider the departure rate from S_1. If there are n lanes in the system and $n\lambda_1 \geq m\mu_1$, then all m servers are busy. Since each server works at rate μ_1, the total departure rate is $m\mu_1$. On the other hand, if $n\lambda_1 < m\mu_1$, then only n servers are busy and the total departure rate is $n\lambda_1$. Since cars emerging from the tollbooth must merge into k lanes in S_2, each of which has arrival rate λ_2, we have

$$k\lambda_2 = n\lambda_1 \implies \lambda_2 = \frac{n\lambda_1}{k}.$$

Final Formula

Based on the discussion above, we get

$$W = W_1 + W_2 = \frac{1}{\dfrac{m\mu_1}{n} - \lambda_1} = \frac{1}{\dfrac{k\nu}{2m\ell} - \dfrac{n\lambda_1}{k}}.$$

Since the problem statement stipulates that under most situations $k = n$, we simplify this formula to

$$W = W_1 + W_2 = \frac{1}{\dfrac{m\mu_1}{n} - \lambda_1} = \frac{1}{\dfrac{n\nu}{2m\ell} - \lambda_1}.$$

The Range of Feasibility

Our model can calculate the optimal number of tollbooths needed only if the denominators for both W_1 and W_2 are greater than zero. Therefore,

$$\frac{m}{n}\,\mu_1 > \lambda_1 \quad \text{and} \quad \frac{n}{m}\frac{\nu}{2\ell} > \lambda_1.$$

Hence the feasible range for the number of tollbooths is

$$\left(\frac{\lambda_1 n}{\mu_1}, \frac{n\nu}{2\lambda_1 \ell}\right).$$

For a single tollbooth per lane, we set $m = n$; the resulting W is

$$W = W_1 + W_2 = \frac{1}{\mu_1 - \lambda_1} = \frac{1}{\frac{\nu}{2\ell} - \lambda_1}.$$

The model is still a system of two queues. Though the merge factor n/m is diminished, the cars must still catch up to traffic speed and may have to wait in a queue to do so.

Data Analysis

We implement our algorithm for W in Matlab using $n = 3$, 4, 5, and 6, corresponding to most tollways. We vary λ_1 from 0.5 to 5 cars/minute for light traffic, from 5 to 10 cars/minute for medium traffic, and from 10 to 15 cars/minute for heavy traffic. We establish the range of feasibility for m for each traffic situation. We then determine the number that gives minimal W.

Parameter Values

We set $\mu_1 = 11$ cars/min for the light and medium traffic; we set $\mu_1 = 7.5$ cars/min for heavy traffic, to account for the service time plus the reaction time of the cars waiting in queue.

To determine μ_2, we set $\nu = 60$ mph for light traffic situations, since most heavily trafficked tollways have speed limits between 50 and 70 mph. We set $\nu = 46$ mph for medium traffic and $\nu = 32$ mph for heavy traffic. The average car length is between 3.5 and 5.5 m [Edwards and Hamson 1990], hence we set car length in our model to 4 m. We set the safety distance to 20 m for light traffic, 14 m for medium traffic, and 8 m for heavy traffic. The optimal number of tollbooths for the different levels of traffic and numbers of highway lanes are shown in **Table 2**.

Table 2.
Optimal numbers of tollbooths.

Traffic	Number of lanes			
	3	4	5	6
Light	5	7	9	10
Medium	5	9	9	11
Heavy	9	9	11	13

Regardless of the traffic level, the optimal number of tollbooths is always greater than the number of highway lanes. However, for light traffic, the difference between the average wait for optimal number of tollbooths vs. the average wait for $m = n$ is only about 2 s.

For medium traffic, though, the differences ($\approx$15 s) are large enough to conclude that having extra tollbooths would be a wise decision.

For heavy traffic, setting single tollbooth per lane would result in infinite waiting queues for all situations examined.

Detailed Analysis of a Six-Lane Tollway

We conduct a detailed study for a six-lane tollway. The general trends observed for this dataset are typical for any number of lanes. We generate plots for the three traffic levels with number m of tollbooths as independent variable and W as dependent variable. We keep λ constant for each curve; hence we produce a set of level curves that show the optimal value for m based on the λs.

As the traffic gets heavier, the region of feasibility for m gets smaller. This is because having too few tollbooths causes an infinite waiting time at S_1, while having too many tollbooths causes an infinite wait at S_2, due to the influx of cars processed in S_1.

For light traffic, the difference in W_{ave} for $m = 5$ and $m = 18$ is merely 2 s. For medium traffic, a shift from the optimal number $m = 10$ causes a more dramatic increase in the time spent in the network. For heavy traffic, the range of feasibility reduces to a small region centered around the optimal number $m = 13$—namely 12, 13, or 14 (**Figure 2**). The onset of heavy traffic both before and after the tollbooth excludes more extreme values of m from the feasible range. The beauty of the results is that the optimal number of tollbooths is the same for varying arrival rates.

Conclusion

It is better to have more than one tollbooth per lane. But having too many tollbooths per lane is just as bad. We recommend that for frequently traveled roads, the number of tollbooths available should be the maximum of all the optimal tollbooth numbers generated by our algorithm. The number of booths open can then vary for different traffic flows during the day.

For toll roads that usually have light traffic, having a single tollbooth per lane reduces the cost of building and running the toll plaza; reduction in waiting time does not justify more tollbooths.

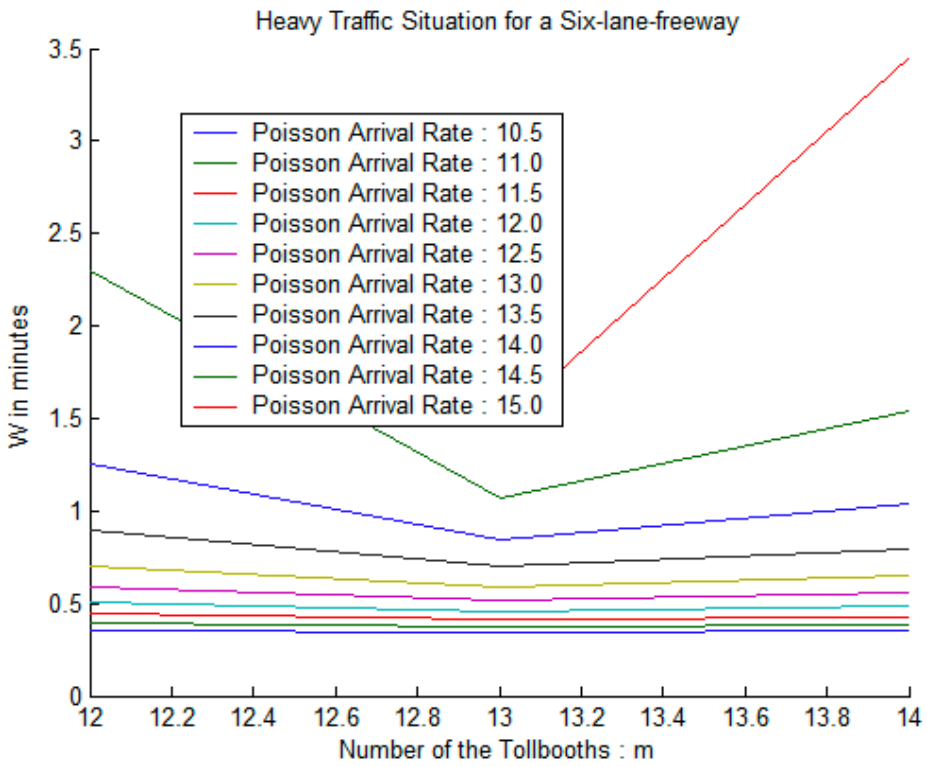

Figure 2. Wait vs. number of tollbooths for heavy traffic on a 6-lane tollway, for various arrival rates.

Strengths of Our Model

- Our model withstands many variations in parameters.
- Given reasonable values for the parameters, the algorithm generates realistic results for the optimal number of tollbooths.
- When we vary within the range of a specific traffic situation, the optimal solution is consistent for each in each situation.
- The optimal number of tollbooths differs among traffic levels, reflecting the fact that varying the number of tollbooths has a significant impact on waiting time.
- The algorithm, though rich in theory, is very easy to implement and test.

Weaknesses

- We assume that the arrival rate is less than the service rate at each system. In the long run, this assumption must hold in order to avoid infinite queues; but there can be intervals during when arrivals overwhelm the service rate.

Hence, though the average waiting time for the optimal solution may be small, the maximum waiting time for some cars may be rather large.

- Our model's range of feasibility is limited by the rates at which the cars are served at the two systems.
- Our model predicts the optimal tollbooth numbers based on the minimal time, but this may not be the most cost-effective solution.
- We don't incorporate the electronic payment passes that many toll systems use to minimize waiting time.

References

Edwards, Dilwyn, and Mike Hamson. 1990. *Guide to Mathematical Modeling*. Boca Raton, FL: CRC Press.

Pitman, Jim. 1993. *Probability*. New York: Springer-Verlag.

Ross, Sheldon M. 2003. *Introduction to Probability Models*. San Diego, CA: Academic Press.

Azra Panjwani, Jim Pitman (advisor), HuanHuan Qi, and Yang Liu.