DOUBLE-SIMULATION METHODOLOGY TO DETERMINE THE NUMBER OF SERVERS IN SITUATIONS WITH PEAKS

Javier Otamendi
Manuel Cano Espinosa
TYPSA – Técnica y Proyectos, S.A.
Plaza del Liceo, 3
28043 Madrid, Spain
E-mail: jotamendi@typsa.es

KEYWORDS
Traffic simulation, queueing systems, system design, flows with peaks, discrete event modelling.

ABSTRACT

A simulation model that represents a multichannel situation with a time-dependent input flow distribution is presented. The objective of the study of the system is to assess its performance under different configurations of the service stations needed to handle the peaks. A methodology is developed based on the execution of a simulation model in two steps. First, to look for the satisficing alternatives, a set of constant flow simulations is run. Second, to obtain the optimum alternative, the model is run with real flows. The presentation is based on standard queueing theory terminology and includes an example of the design of a paytoll booth in a highway in which the incoming traffic is highly seasonal.

INTRODUCTION

A system is a combination of units that look for service. An airplane (unit) trying to take-off in a given runway (service station), a vehicle trying to refuel at a gas station, a customer trying to make a deposit… They are all examples of real systems, in which if the service station does not have enough capacity, queues are formed with the corresponding hassle for the customers. Therefore, when designing the system (Rubinstein 1986), there must be a compromise between cost (or built capacity) and customer service (waiting times in queues).

The decision is specially difficult when the arrival of units is not smooth or constant but following peaks that depend on time. More planes take off early in the morning or late in the afternoon than in the middle hours of the day, more vehicles stop at a gas station early morning or late afternoon, banks usually have a more stable arrival of customers…

The system that is the focus of this article is a paytoll station in the highway at the entrance of a big city. The arrival of vehicles is clearly seasonal, with tremendous peaks almost every Sunday afternoon and the day of the return from a regional holiday. Two or three-hour peaks in which the hourly arrival rate is sometimes 100 times more than on a week day. A decision has to be taken about the number of booths to install to cover most of the incoming flow but without overdimensioning the system (Gross and Harris 1985).

Queueing models are analytical models that are used to analyse and quantify the performance of a service station. They are analytical models that, for specific combinations of input flow, service times, number of servers and queue disciplines, estimate the queue lengths and times in the system (Taha 1987). The requisite, however, is that the capacity installed must be enough to handle all the incoming flow.

Even if that condition holds, not all the situations might be represented with analytical models (for example, moving from queue to queue, not constant incoming service rate) and other tools that provide for a good abstraction of the system have to be used.

Simulation models are descriptive models that might reflect complicated queue disciplines and any distribution for the input parameters, especially when they vary with time.

A combination of queueing theory and simulation modelling is proposed to analyze the seasonal incoming flow at the service stations and to design the system to reach a satisfactory compromise between cost and customer service.

In order to determine the capacity of the system to be installed, it is necessary to follow a reasonable methodology that searches through the available alternatives and selects the satisficing one. This methodology has to be based on the fact that both queueing and simulation models are descriptive models and not optimization models, since they are not used to optimize a given situation but to describe it. That means that, for a given set of input parameters, the model is executed and the performance estimated. Since the objective of the study is to search for a satisficing alternative, then, the only way is to vary the values of the input parameters and estimate the objective function for each combination, with a

Proceedings 16th European Simulation Symposium
György Lipovszki, István Molnár © SCS Press, 2004
ISBN 1-56555-286-5(book) / ISBN 1-84233-106-x(CD)

posterior step to compare and select the alternative with the best values for the set of input parameters.

This search process is depicted in the following figure, which includes a four-step procedure:

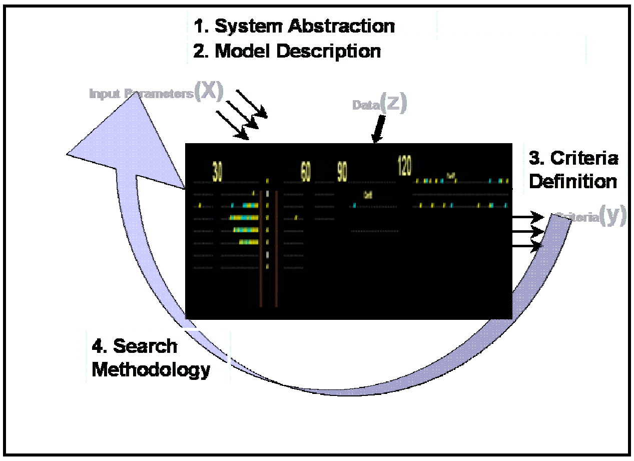


Figure 1. Search Process

The first and second steps are used to describe the actual situation, including the possible input variables. The third is to develop a suitable objective function to compare and select among alternatives. The final step is the iteration process in which values for the input parameters are changed.

DESCRIPTION OF THE SYSTEM/MODEL

Let us use this section to describe the system under consideration. It is just a multichannel service station in which customers arrive and ask for service.

A common way used to describe a queueing system is Kendall's notation in which six parameters must be specified (Taha 1987):

(a/b/c):(d/e/f)

where

a: input flow distribution with average λ
b: service time distribution with average μ
c: number of service stations
d: queue discipline (FIFO, LIFO, …)
e: available queue length
f: maximum number of arrivals.

Input Flow

The main characteristic of peaks is that the input flow varies with time. There are periods in which the number of customers that arrive is significantly greater than the average as to determine that the distribution is not the same along the time horizon of the study.

Let us define, then, a distribution function f(x), measured in arrivals per unit time, and whose expected value is λ. Its cumulative function is F(x). Figure 2 shows how the real input flows along time might be converted into a probability distribution function.

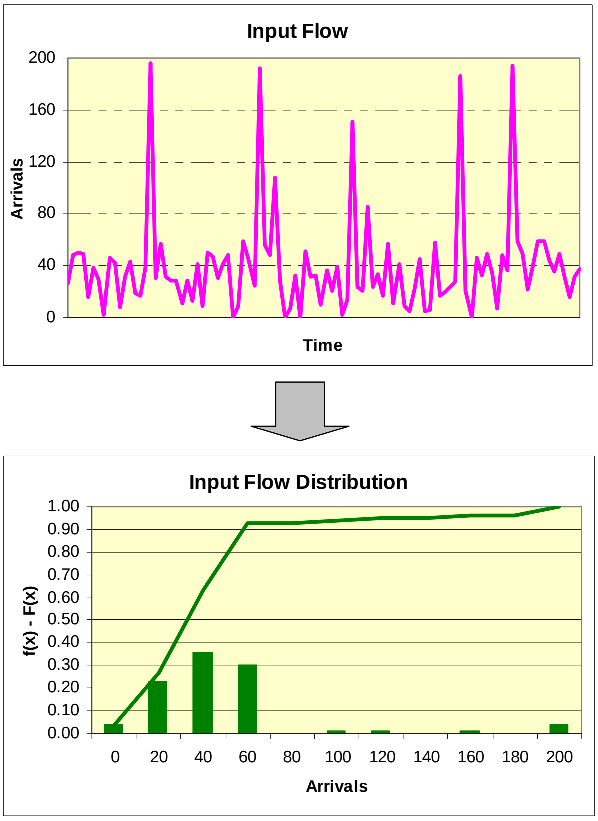


Figure 2: From Real Flows to Input Flow Distribution

The average is in this case meaningless since it is not constant with time. What matters is the maximum number that can arrive per unit time.

Service Time

Let us define throughput as the number of customers that a single station can serve per unit time. This value is usually not dependent on time, since the service rate is the same regardless of the number of customers waiting or the number of servers available. The distribution is usually multimodal, since the customers can be differentiated in types, each with a given distribution. Let's call this distribution f(y), its expected value μ and its cumulative distribution F(y). Figure 3 shows an example of a distribution of this type:

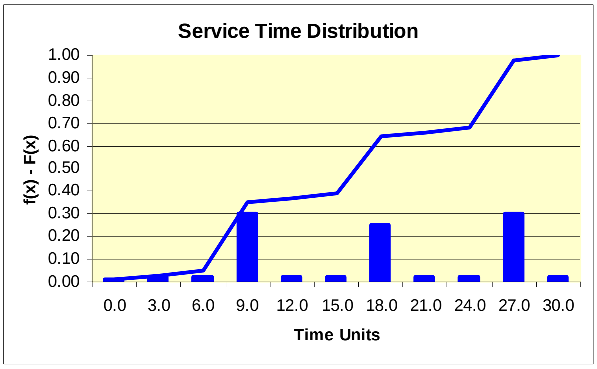


Figure 3: Service Time Distribution

Service Stations

The service stations are set in parallel, each with its own queue. The space in front of the stations is arbitrarily long. The number of service stations are S, the quantity that needs to be calculated. Notice that usually this number has an upper bound, S', which depends on the space available and on budget.

The number of servers directly determines the capacity of the system, CAP, or the average number of customers that the whole set of stations can serve per unit time, which is calculated as:

$$CAP = \mu * S$$

It is also the maximum input flow that the system can handle without the queues growing infinitely.

Figure 4 shows this relationship in graphical form:

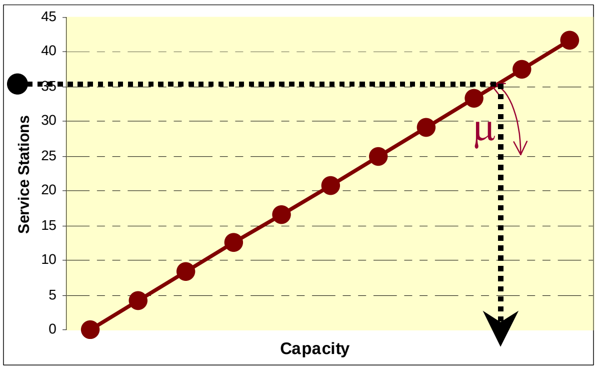


Figure 4. Capacity as a Function of S

For any given S, capacity can be calculated. The slope of the line depends on the value of the mean service time μ.

Queue Discipline

The customers are served on a first-come first-served basis in each of the servers. However, changing queues is allowed. The vehicles cannot leave the system.

Available queue length

It is the available space for queues to be formed. In the situation under study, there is no limit since the incoming flow waits for service even in the highway.

Size of Source

It is the total number of possible clients that access the service stations. In the situation in hand, the total number of input customers is known in an hourly basis.

MEASURES OF PERFORMANCE

In this section, the most important measures of performance are defined, quantified and related mathematically. Queueing models are used to calculate queue lengths and times in the system, as well as percent utilization. They are all related in terms of λ, μ and S. In this case, let us take the queue length as the driver and percent utilization as a summary measure. Also, a cost criteria and idea of flow coverage are included.

Cost

Let's assume that the cost increases linearly with the number of servers:

$$COST = S * \$$$

where $ is the cost per service station.

Coverage

The idea of flow coverage specifically applies to situations with peaks. Analytical queuing models usually require that the input flow is less than the capacity for the system to be stationary or stable.

In situations with peaks is economically infeasible to design the system to cover all situations. The design capacity will cover most of the income flows but not all the peak flows. Even in these busy systems, sometimes it is desired to install less capacity to control the flow intensity downstream (Huang and Huang 2002).

Let's define then coverage (COV) as the percentage of income flow that corresponds with the capacity of the system:

$$F(x=CAP) = COV$$

In Figure 5, the actual coverage, that is, the one that corresponds to the capacity of the system, is calculated using the input flow distribution.

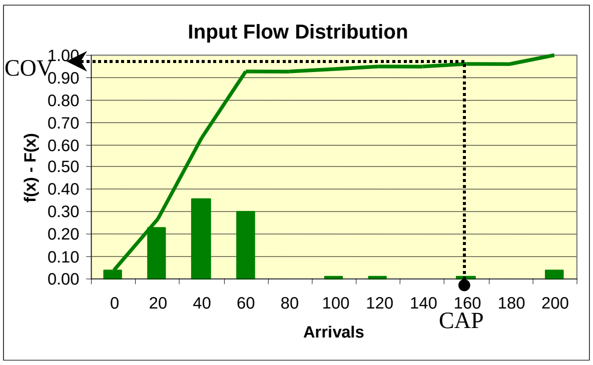


Figure 5: From Capacity to Coverage

The relationship between CAP and COV is then crucial in the design of the system. It is obvious that the higher the desired coverage the higher the capacity required. Also, it must be mentioned that when the coverage is already high, small increases in coverage require a large increase in capacity.

Since CAP depends directly on S, there is also a direct relationship between the number of servers and the coverage, which is depicted in Figure 6.

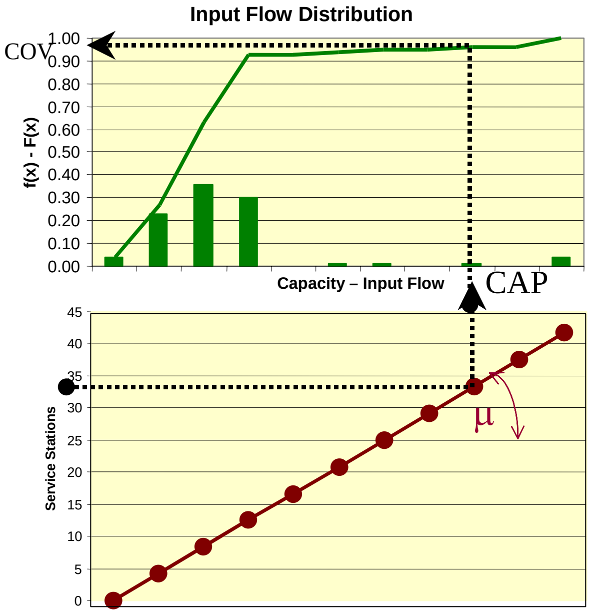


Figure 6. From Service Stations to Coverage

The number of servers sets the capacity, which is equivalent to input flow. The corresponding coverage might then be calculated.

The lack of coverage (1-COV) is also an important measure since it is the percentage of time in which customers arrive at a higher rate than the service rate and large queues will be formed. That percentage corresponds to the peaks.

Utilization

Utilization, %UTIL, is the percentage of time the servers are attending customers. It also might be defined as the ratio between what arrives to the system and what might be served:

$$\%UTIL = \lambda / CAP$$

where λ has already been defined as the average input flow and CAP as the average capacity of the system.

Therefore, the utilization is to be calculated as:

$$\%UTIL = \lambda / CAP = \lambda / (\mu * s) \qquad (1)$$

Then, coverage might also be defined as the percentage of time in which the queues are manageable or the percentage of time in which %UTIL $\leq$ 100%.

Queue Length

The length of the waiting queues (LQ) is going to vary significantly between the normal hours of operation and the peak hours. The relationship that must be studied is between Utilization and LQ, which will be something similar to the one shown in Figure 7. The graph has been obtained using the M/M/S theoretical queueing model (Taha 1987).

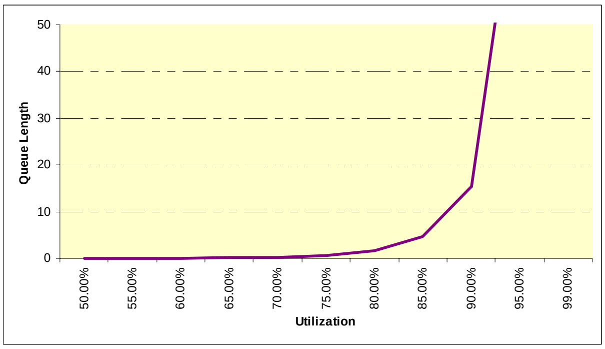


Figure 7. Relationship Between Utilization and Queue Length

If the utilization of the system is low (around 60%), the queues are non-existent, whereas if the utilization is greater than 1, the queues grow to infinity. There is a grey area, that corresponding to a %UTIL close but lower than 1, in which the queues are manageable but usually longer than desired.

The range of the grey area depends on the nature of both the input flow and the service time distributions, as well as the number of service station. An upper limit for the %OCUP must be investigated, as it will definitely influence on the design decision, since it is a function of the permissible queue length.

SEARCH METHODOLOGY

The objective of the methodology is to calculate the optimum number of servers needed to balance the different criteria involved in the design of the system. The more servers are included, the higher the capacity, the higher the coverage, the smaller the queue length, but the higher the cost.

With the graphs included in the previous section, a trial-and-error procedure might be implemented so that for any given value of S, the performance measures are calculated, the alternatives compared against each other and one of them selected.

However, this procedure can be very tedious if the possible number of servers is too large. For that reason, what is proposed is to perform a first step to reduce the number of feasible alternatives to a manageable set.

The idea is to use the performance criteria also as restrictions by setting bounds on their values. The procedure is therefore to work on an optimization mode, and instead of calculating the performance measures for a given value of S, determine the number of servers that correspond to a particular limiting value of each criteria.

What must be a priori provided then is a set of satisficing values on each of the measures that will help do the balancing. Let’s define those limits as:

- BUDGET = Maximum cost that might be invested in servers. Obviously, if money is not a factor, the service station could be built with as many servers as desired and no queues will ever be formed.
- ALQ = Maximum average queue length. It is the satisficing value for the average number of customers waiting for service.
- S%UTIL = Satisficing level for the utilization of the servers. It is defined as a level in which the average queue length is kept below its satisficing limit.
- mCOV = Minimum coverage. It is the minimum value of income flow that must be attended without forming queues.

The search methodology follows a five-step procedure that includes obtaining input data and a double simulation. The result is a range of feasible values for the number of servers needed to satisfy the utilization, the coverage, the queue length and the cost requirements.

Step 1: Obtain data

1a. Input flow distribution
Historic data of flows per unit time is obtained and converted into an absolute frequency distribution, its corresponding relative frequency or probability distribution f(x) and the cumulative probability distribution F(x) (see Figure 1).

1b. Expected value of the service time distribution, μ
Historic data of time spent in the service station per unit is obtained and its average calculated.

Step 2: Determine the admissible input flow

2a. Define minimum coverage
The value of mCOV is defined subjectively.

2b. Calculate the minimum coverage flow
The input flow value corresponding to that coverage value might then be calculated as:

$$F\_mCOV = F^{-1}(mCOV)$$

Figure 8 shows the calculation in graphical form:

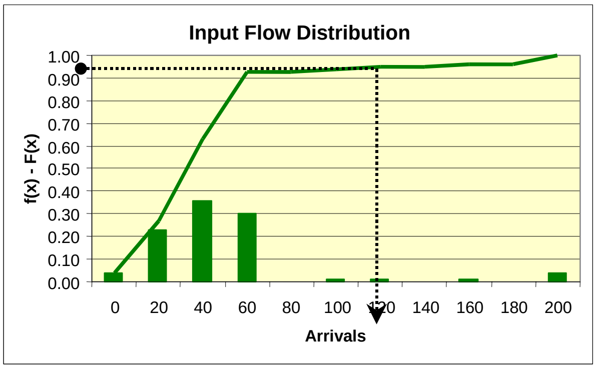


Figure 8. Calculation of coverage flow

2c. Calculate the admissible input flow
The coverage flow might be denominated also as the admissible input flow F_ADM:

$$F\_ADM = F\_mCOV$$

It is then the absolute minimum flow that must be attended without forming queues.

Step 3: Calculate the feasible range on the number of servers

A feasible range for the number of servers can now be calculated. For the upper bound, S’, the BUDGET constraint is used:

$$S' = BUDGET/\$$$

For the lower bound, ‘S, the value is estimated knowing the minimum coverage and its desirable input flow (calculated in Step 2c) and the satisficing utilization, using Equation (1) as follows:

$$`S = \sup (F\_ADM / (\%OCUP * \mu) \qquad (2)$$

where sup() indicates rounding to the higher integer.

Figure 9 shows the calculation of the lower bound for different values of %OCUP:

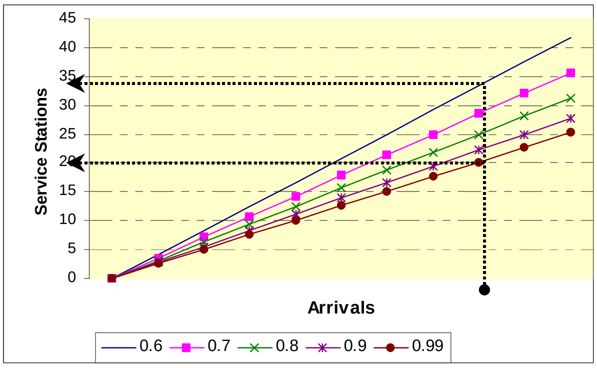


Figure 9. Calculation of ‘S

Several feasible values for S are now remaining, each providing a satisficing compromise between cost and service (queue length). Among these values, the greater the S, the better the service but the higher the cost.

Step 4: Obtain the satisficing occupation level via constant-input simulation.

The satisficing %OCUP has been defined as the one that corresponds with a reasonable queue length, ALQ. This S%OCUP will vary with ALQ, but also with number of servers S and the mean service time μ.

As it has been already demonstrated, in general, if the %UTIL is close to 1, the queues will be too long, and if %UTIL is close to 0, there will be no queues. But the relationship in the rest of the range cannot be quantified easily. A value of 60% is considered as a lower bound below which no queues are developed.

Here is where simulation or queueing models come into play. If an analytical model corresponds to the situation in hand, which is not usually the case, queueing models are to be used to calculate the average queue length. If the situation cannot be analytically modelled, a simulation study can be performed. This is the case in which the input flow distribution presents heavy peaks and the service distribution is multimodal.

In this step, the simulation model is to be used to evaluate the mean queue length under varying conditions of utilization. For that reason, it is not worthwhile to execute the model using the real flow with peaks, but with a constant average input flow, λ. If the simulation model is run for different values of λ and S, but for the real values of the service distribution, a feel for the value of the average queue length can be obtained:

$$LQ = f(\%OCUP) = f(\lambda,S) \text{ for a given } \mu$$

A satisficing value of S%OCUP might then be estimated. This value will correspond to the maximum %OCUP that satisfies the ALQ requirement:

$$S\%OCUP = \max \%OCUP \mid LQ \leq ALQ$$

With this value, ‘S is to be more closely estimated.

Step 5: Compare the feasible alternatives via real-flow simulation

But obviously, the study cannot forget about the actual flows, that is, the evaluation of the flow with respect to time. It’s then the appropriate moment to use the simulation model for the second time.

With real flows, the simulation model can be executed for the satisficing values of the variable S and the relationship between cost and service quantified. The influence of the peaks is then dynamically studied and the final decision taken accordingly.

Graphical tools for calculating the feasible range of servers

The sequential procedure taken in the first three steps might be easily performed with the use of a pair of simple graphs. The double-graph tool is shown in Figure 10.

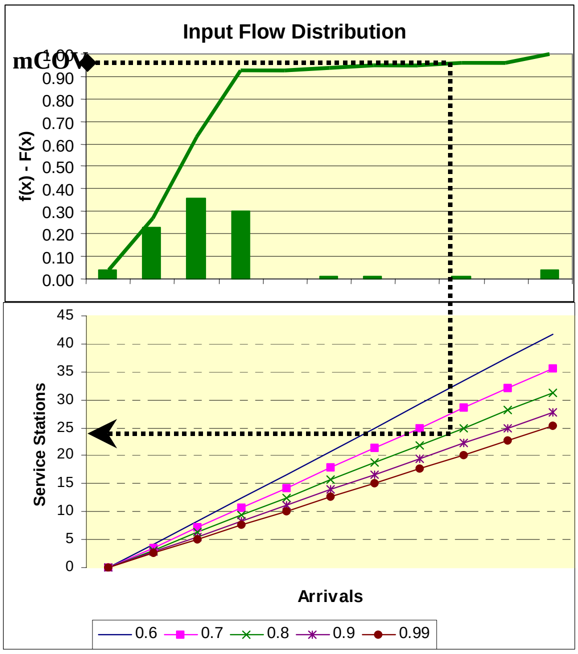


Figure10. Double-Graph Analysis

The first one uses the minimum coverage as the input value to calculate the admissible input flow (Steps 1-2). This value is used in the second graph with the calculated satisficing occupation value (Step 4) to provide a lower bound for the number of servers (Step 3).

An equivalent single-graph tool is shown in Figure 11. The results obtained are the same, but information concerning the allowable input flow is lost.

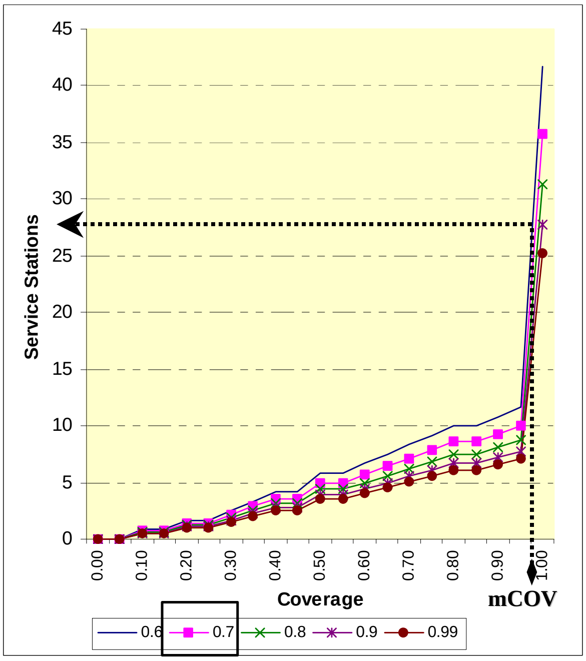


Figure 11. Single-Graph Analysis

The shape of the graph, and therefore the final decision, depends directly on the input flow distribution and the number of classes that have been used to develop its frequency distribution.

PAYTOLL EXAMPLE

One of the real-life systems in which significant queue lengths are formed caused by an input flow distribution with large peaks is the paytoll system in highways. The design of these systems is not easy since a large percentage of the time the input flow is not important, whereas a small percentage of time the input flow is tremendous. In a country like Spain, there exists a huge agglomeration of people going in and out of the big cities during a holiday period.

The difference in flow between a high and a low period makes it economically impossible to design the paytoll system to cover all the possible incoming flows. A design tool is necessary to balance the counteracting effects of cost and service while representing the busy situation over long periods of time.

What follows is a description of the analysis that was performed to analyze one of these paytolls, using the methodology that is the focus of this article. The input flow and service distributions have been altered to preserve proprietary information.

Step 1a: Obtain the input flow distribution

The flow was obtained for a period of 6 months, counting the number of vehicles that had entered the highway per hour. Figure 12 is the histogram that has been obtained after treating the data.

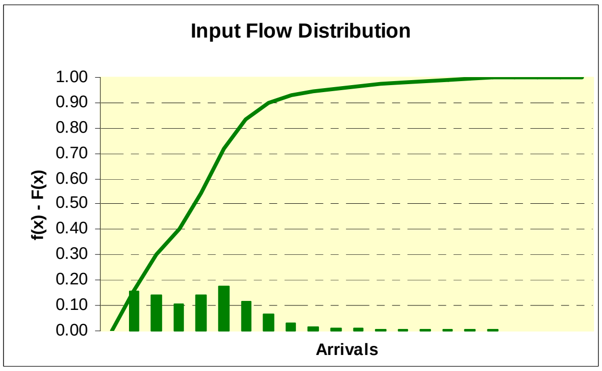


Figure 12. Input Flow Distribution

About 1/3 of the possible input values account for 90% of the data. Higher cumulative frequency is only achieved then if the arrival rate is increased significantly, showing that few but high peaks appear. In fact, the peaks correspond to Sundays when people return to the big city from the mountain or the beach.

Step 1b: Calculate the expected value of the service time distribution, μ

Historic data of time spent in the service station per unit is obtained and represented in Figure 13. The distribution shows two modes, corresponding to payment using credit cards or cash.

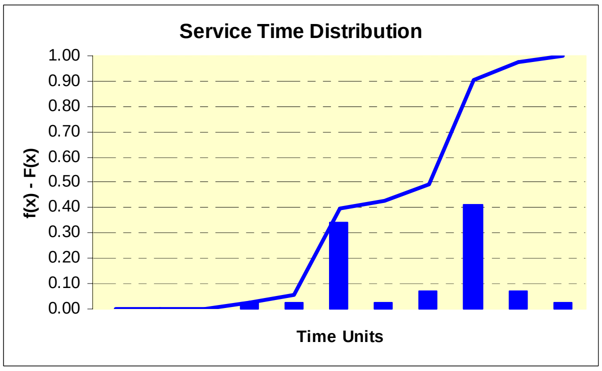


Figure 13. Service Time Distribution

Step 2 & 3: Calculate the feasible range of number of servers

The value of mCOV is first defined subjectively to 98%.

The use of the double-graph analytical tool is then used to determine the feasible lower bound on the number of service station. Figure 14 shows the process.

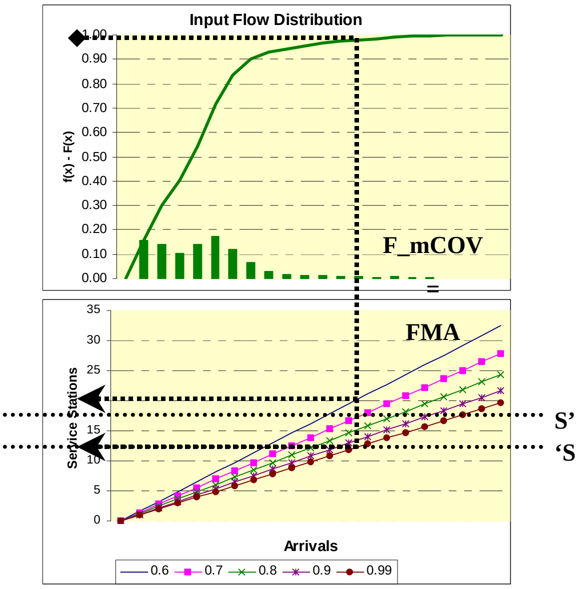


Figure 14. Calculation of the Feasible Limits

For a subjective mCOV value of 98%, the corresponding minimum coverage flow F_mCOV is calculated, value which is also equal to the admissible flow FMA.

If no queues are to be formed (60% occupation) there is a need for 20 service stations, whereas if the occupation strived for is almost 100%, with 12 servers is enough, this level becoming the lower limit 'S.

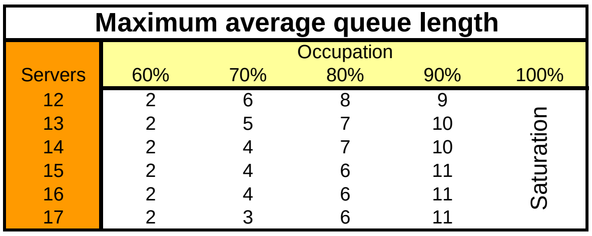
The budget however imposes an upper limit S' = 17 service stations. Therefore,

$$12 \leq S \leq 17$$

Step 4: Obtain the satisficing occupation level via constant-input simulation.

The aim of this important step is to determine the admissible queue length for the admissible coverage flow.

A set of constant flow simulation runs is performed, varying the number of servers in the feasible range between 12 and 17 and the occupation level between 60% and 100% in increments of 10%, measuring in each case the average queue length. Figure 15 summarizes the results.



Maximum average queue length					
	Occupation				
Servers	60%	70%	80%	90%	100%
12	2	6	8	9	Saturation
13	2	5	7	10	
14	2	4	7	10	
15	2	4	6	11	
16	2	4	6	11	
17	2	3	6	11	

Figure 15. Queue length versus Occupation

The results show that saturation is important once the occupation is above 90% and that the queue length decreases slightly as the number of servers increases, except for an occupation of 90%. The reason in this last case is that the total number of vehicles or admissible input flow becomes too large (heavy traffic conditions) when the number of servers grows (Huang and Huang 2002).

Management then determined also that 8 cars had to be the acceptable maximum and that 6 was probably the target, fixing therefore the occupation level between 70% and 80%.

Translating these results into the single-graph tool (Figure 16), the range of the feasible number of servers can be narrowed.

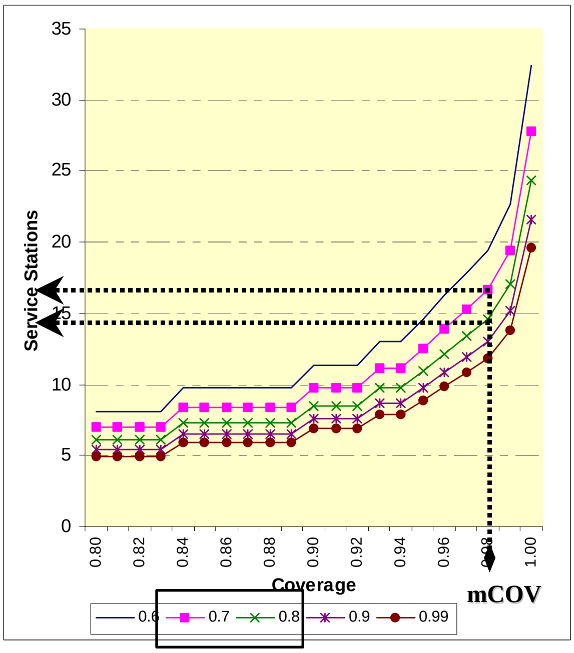


Figure 16. Calculation of the Tight Feasible Limits

The new feasible range is between 15 and 17 service stations:

$$15 \leq S \leq 17$$

Step 5: Compare the feasible alternatives via real-flow simulation

The simulation model with real flows over a six-month period is run for the three available alternatives, measuring queue length.

For the alternative with 15 service stations, there are 17 days in which the system is collapsed when the users return from holidays (Figure 17). Almost every Sunday the queue length is larger than the desired level. The rest of the days, the queue is basically non-existent.

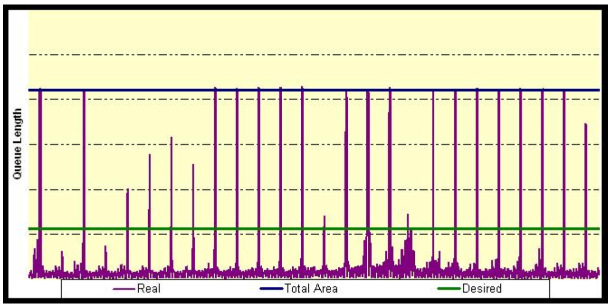


Figure 17. Simulation with Real Traffic and S = 15

In Figures 17, 18 and 19, the total available paytoll area is also included to show that in certain days the queues that are formed go beyond the paytoll area and even collapse the highway.

For the alternative with 16 service stations (Figure 18), the performance is obviously improved, with only 10 days collapsing the highway.

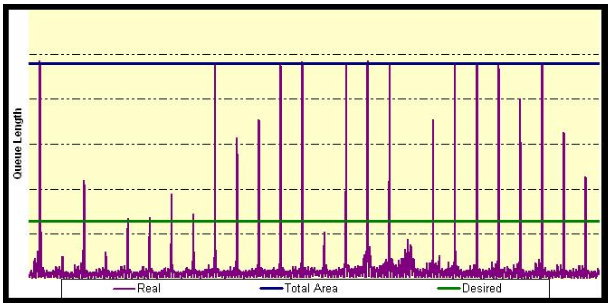


Figure 18. Simulation with Real Traffic and S = 16

Finally, with 17 service stations (Figure 19), only 1 day shows problems and 16 Sundays are above the desired level.

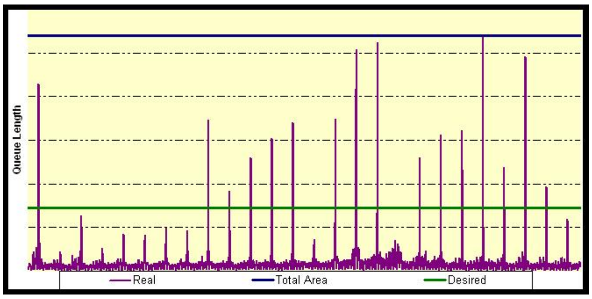


Figure 19. Simulation with Real Traffic and S = 17

At this point, management decides to eliminate the alternative with 15 servers. Besides the sub par performance, it presents side effects like blockage at some entry points into the highway several kilometres back.

The final decision was then between an increase in performance (17 booths) and the save in money (16 booths).

SUMMARY AND CONCLUSIONS

A methodology is presented to design systems that suffer short periods of very high peaks based on both queueing theory and simulation modelling. The proposed methodology follows a five-step procedure that includes obtaining the input flow and the service time distributions, a simulation model that is run both with constant and real flows and a double-graph summary tool.

This double use of simulation has been proven very successful since the time spent in running the model was very small compared to the time spent in collecting the data and developing the model. Once the model was validated, it was used for a thorough experimentation period that led to a strong-based decision.

A real-life example of a paytoll system in a highway is presented to validate the methodology. The limits on the number of booths are set quickly and the simulation runs help to make the final decision.

Therefore, the combination of experimentation via simulation models with quantitative queueing analysis has been proven as an interesting tool to treat situations with peaks that incorporates both management input and intensive collection of data.

REFERENCES

Gross, D. and C. Harris. 1985. *Fundamentals of Queueing Theory*. John Wiley and Sons, New York.

Huang, D. and W. Huang. 2002. “The effects of tollbooths on highway traffic.” *Physica A 312*, No.3-4, 587-608.

Rubinstein, R. 1986. *Monte Carlo Optimization, Simulation, and Sensitivity of Queueing Networks*. John Wiley and Sons, New York.

Taha, H.A. 1987. *Operations Research*. Macmillan, New York.

AUTHOR BIOGRAPHIES

JAVIER OTAMENDI received the B.S. and M.S. degrees in Industrial Engineering at Oklahoma State University, where he developed his interests in Simulation and Total Quality Management. Back in his home country of Spain, he received a B.S. in Business Administration and a Ph.D. in Industrial Engineering. He is currently a simulation and statistics consultant and university professor. His e-mail address is: jotamendi@typsa.es

MANUEL CANO ESPINOSA received the Civil Engineering degree from the Universidad Politécnica de Madrid in 1983. He has worked as a supervising director in road and highway design projects, both in Spain and in Europe and Latin America. He is now the Project Manager of the Road Construction Division at the consulting firm Técnica y Proyectos S.A. (TYPSA). His e-mail address is: mcano@typsa.es