Article ECORFAN Journal-Republic of El Salvador
December 2019 Vol.5 No.9 1-12

Optimization of import transfers from a customs bonded warehouse using a server model

Optimización de transferencias de importación de un recinto fiscalizado aplicando un modelo de servidores

NUÑEZ-PEREZ, F. A. *† & ESCOTO-SOTELO, E. A.

Universidad Politécnica de Lázaro Cárdenas, 60950, Nitro 40, Nuevo INFONAVIT, Lázaro Cárdenas, Mich.

ID 1st Author: *F. A. Nunez-Perez* / **CVU CONACYT ID:** 164252

ID 1st Coauthor: *E. A. Escoto-Sotelo* / **CVU CONACYT ID:** 390313

DOI: 10.35429/EJRS.2019.9.5.1.12 Received July 28, 2019; Accepted December 15, 2019

Abstract

It is currently necessary to implement proposals that reduce service level times in the area of customs control. For this it is necessary to carry out a study of mathematical models that adapt to the reduction of possible problems. Once an appropriate predictive pattern is found, it is necessary to carry out an implementation to find the best option allowing to obtain a continuous flow of service, achieving optimal scaling. But above all achieving a better control in all the processes developed with better fluidity in the input and output modules and in the operational area.

Optimization, Mathematical models, Customs

Resumen

En la actualidad es necesario implementar propuestas que disminuyan los tiempos de nivel de servicio en el área de control de aduanas. Para ello es necesario realizar, un estudio de modelos matemáticos que se adapten al abatimiento de posibles problemas. Una vez de encontrar un patrón predictivo apropiado, es necesario llevar a cabo una implementación para encontrar la mejor opción permitiendo obtener un flujo continuo de servicio, alcanzando un óptimo escalamiento. Pero sobre todo alcanzando un mejor control en todos los procesos desarrollados con una mejor fluidez en los módulos de entrada como de salida y en área operativa.

Optimización, Modelos matemáticos, Aduanas

Citation: NUÑEZ-PEREZ, F. A. & ESCOTO-SOTELO, E. A. Optimization of import transfers from a customs bonded warehouse using a server model. ECORFAN Journal-Republic of El Salvador. 2019. 5-9: 1-12.

* Correspondence to Author (email: Phd_paco@hotmail.com)
† Researcher contributing first author.

Introduction

The inspected enclosures are concessions granted by the tax administration service, with the objective of having an adequate service that involves the handling, storage and custody of merchandise. The case study reported in this project was carried out in the audited premises 199 and 221 according to appendix 6 of the customs law. Attached to customs law, site 221 provides the following services: Loading / unloading of merchandise from ship to ship side or vice versa. Shipment of merchandise from ship side to storage area. Delivery / receipt of merchandise from storage area by means of transport or vice versa (rail or truck), examinations of goods (prior), deconsolidation, consolidation, labeling and conservation of goods. With the knowledge of the cause of the multiple services listed, a search for the optimization of the audited area was performed, analyzing the following optimization models:

Optimization models

Queue theory: is responsible for the mathematical analysis of the phenomena of waiting lines or queues. These types of models are frequently presented when a service is requested by a series of clients and both the service and the customers are probabilistic. The study of the waiting lines tries to quantify the phenomenon of waiting in queues, through representative measures of efficiency, such as the average length of the queue, the average waiting time in it, as well as the average use of the facilities.

Elements of a queue model: the main actors in a queue situation are the client and the server. Clients arrive at an installation (service) from a source. Upon arrival, a customer can be serviced immediately or wait in a queue if the facility is busy. When an installation completes a service, it automatically "pulls" a customer who is waiting in the queue, if any. From the point of view of queue analysis, the arrival of customers is represented by the time between arrivals (time between successive arrivals), and the service is measured by the service time per customer. Usually, the time between arrivals and service are probabilistic or deterministic.

The size of the tail plays a role in queue analysis. It can be finite, for all practical purposes, infinite. Queue discipline represents the order in which customers are selected in a queue. This factor is of great importance in the analysis of queue models. Having the following disciplines.

A) The first to arrive is the first to be attended (the most common).

B) The last to arrive is the first to be served.

C) The service in random order.

D) Select customers from the queue, based on some order of priority.

Queue behavior plays a role in the analysis of waiting lines. Customers can switch from a longer to a shorter queue to reduce the waiting time, they can stop queuing due to the long anticipated delay, or get out of a queue because they have been waiting too long.

The service installation design may include parallel servers. They can also be arranged in series or arranged in a network. The source from which customers are generated can be finite or infinite. A finite source limits the number of customers that arrive. An infinite source is, for all practical purposes, forever abundant.

A queue system is specified by six main features:

- The type of distribution of tickets or arrivals (time between arrivals).

- The type of distribution of exits or withdrawals (service time).

- The service channels.

- The discipline of service.

- The maximum number of clients allowed in the system.

- The source or population.

The objectives of queue theory consist of:

- Identify the optimal level of system capacity that minimizes its overall cost.

- Evaluate the impact that the possible alternatives for modifying the capacity of the system would have on its total cost.

- Establish a balanced ("optimal") balance between quantitative considerations of costs and qualitative considerations of service.

- Pay attention to the time spent in the system or in the queue: the "patience" of customers depends on the type of specific service considered and that can cause a customer to "leave" the system.

Role of exponential distribution

In most queuing situations, arrivals occur randomly. Randomness means that the occurrence of an event (for example, the arrival of a customer or the termination of a service) is independent of the time elapsed since the occurrence of the last event. Random times between arrivals and service are quantitatively described by means of an exponential distribution, which is defined as Eq. (1).

$$f(t) = \sum_{k=0}^{n} \lambda e^{-\lambda t}, t > 0 \tag{1}$$

For exponential distribution Eq. (2)

$$E\ \{t\} = 1/\lambda \tag{2}$$

$$P\ \{t \ \leq \ T\} = \int_0^T \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda T}$$

The definition of E (t) shows that it is the rate per unit of time at which events (arrivals or departures) are generated. The exponential distribution describes a totally random phenomenon. For example, if the time is now 8:20 AM. Whereas the last arrival was at 8:02 AM. The probability that the next arrival will occur at 8:29 is a function only of the interval from 8:20 to 8:29, and is totally independent of the time that has elapsed since the occurrence of the last event (8:02 a 8:20 AM). The totally random property of the exponential is known as forgetfulness or lack of memory. Since f (t) is the exponential distribution of time t, between successive events (arrivals), if S is the interval from the occurrence of the last event, then the forgetfulness property implies that:

$$P\{t > T + S | t > S\} = P\{t > T\} \tag{3}$$

To verify this result, we observe that for the exponential with mean 1 / λ, Eq. (4).

$$P\{t > Y\} = 1 - P\{t < Y\} = e^{-\lambda Y} \tag{4}$$

Therefore:

$$P\{t > T + S | t > S\} = \frac{P\{t>T+S, t>S\}}{P\{t>S\}} = \frac{P\{t>T+S\}}{P\{t>S\}}$$

$$= \frac{e^{-\lambda(T+S)}}{e^{-\lambda S}} = e^{-\lambda T}$$
$$= P\{t > T\}$$

Pure birth and death models (relationship between exponential and poisson distribution)

This section presents two queue situations, the pure birth model in which only arrivals occur, and the pure death model in which only exits occur. An example of the pure birth model is the creation of birth certificates of newborn babies. The model of pure death can be demonstrated through the random withdrawal of an item in existence in a store. The exponential distribution is used to describe the time between arrivals in the pure birth model and the time between exits in the pure death model. A byproduct of the development of the two models is to demonstrate the close relationship between the exponential distribution and that of Poisson, in the sense that one distribution automatically defines the other.

Pure birth model

Only arrivals occur.

Define: P0 (t): probability that no arrivals will occur during a period of time t. Since the time between arrivals is exponential and the arrival rate is λ clients per time unit, then: Eq. (5)

$$P0(t) = P\{\text{ time between arrivals} \geq t\} \tag{5}$$

$$= 1 - P\{time\ between\ arrivals\ \ \leq t\}$$

$$= 1 - \left(1 - e^{-\lambda t}\right)$$

For a sufficiently small time interval *h* > 0, we have: Eq. (6)

$$P0(h) = e^{-\lambda h} = 1 - \lambda h + \frac{(\lambda h)^2}{2!} - \cdots = 1 - \lambda h + 0(h^2) \quad (6)$$

The exponential distribution is based on the assumption that during h> 0, when much an event (arrival) may occur. Therefore, as h → 0, Eq. (7)

$$P1(h) = 1 - P0(h) \approx \lambda h \quad (7)$$

This result shows that the probability of an arrival during h is directly proportional to h, with the arrival rate, λ, as a constant of proportionality. To derive the distribution of the number of arrivals during a period t when the time between arrivals is exponential with average 1 / λ, define:

$Pn(t) =$ Probability of n arrivals during t.
For a h> 0 small enough, Eq. (8).

$$Pn\ (\mathrm{t+h}) \approx Pn\ (\mathrm{t})\ (1\text{-}\lambda\mathrm{h}) + Pn\ \text{-}1\ (\mathrm{t})\lambda h,\ n > 0 \quad (8)$$

$$P\mathrm{o}\ (\mathrm{t+h}) \approx P0\ (t)\ (1\text{-}\ \lambda h), \qquad n=0$$

In the first equation there will be n arrivals during t + h if there are n arrivals during t and no arrival during h, or n - 1 arrivals during t and one arrival during h. not all other combinations are allowed because, according to the exponential distribution, at most there may be an arrival during a very small period h. The law of the product of probabilities is applicable to the right side of the equation because arrivals are independent. As for the second equation, during t + h there can be 0 arrivals only if there are no arrivals during t and h. Rearranging the terms and taking the limits as h → 0 to obtain the first derivative of Pn (t) with respect to t, we have:

$$P'n(t) = \frac{\lim\limits_{h\to 0} Pn(t+h) - Pn(t)}{h} = -\lambda Pn(t) + \lambda Pn - 1(t), n > 0$$

$$P'0(t) = \frac{\lim\limits_{h\to 0} P0(t+h) - P0(t)}{h} = -\lambda P0(t),\ n = 0$$

The solution of the above differential equations gives Eq. (9)

$$Pn(t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}, n = 0,1,2,\ldots \quad (9)$$

This is a poisson distribution with mean E {n | t} = λt of arrivals during t. The previous result shows that, if the time between arrivals is exponential with average 1 / λ, then the number of arrivals during a specific period t is Poisson with average λt. The opposite also works. The following table summarizes the relationships between the exponential distribution and Poisson, given the arrival rate λ:

Pure death model

In the model of pure death, the system starts with N clients at time 0, with no new arrivals allowed. Departures occur at the rate of m customers per unit of time. To develop the differential equations of the probability Pn (t) that n clients remain after t units of time, we follow the arguments used with the pure birth model. Thus,

$PN(t+h) = PN(t)(1 - \mu h)$
$PN(t+h) = PN(t)(1 - \mu h)$
$Pn(t+h) = Pn(t)(1 - \mu h) + Pn +$
$1(t)\mu h, 0\ < n < N$
$P0(t+h) = P0(t)(1) + P1(t)\mu h$
As h → 0, we get
$P'N(t) = -\mu PN(t)$
$Pn(t+h) = Pn(t)(1 - \mu h) + Pn +$
$1(t)\mu h, 0\ < n < N$
$P'0(t) = \mu P1(t)$

The solution of these equations gives the following truncated Poisson distribution Eq (10):

$$\mathrm{Pn(t)} = \frac{(\mu t)^{N-n} e^{-\mu t}}{(N-n)!}, n = 1,2,\ldots, N \quad (10)$$

$$\mathrm{P0(t)} = \sum_{n-1}^{N} Pn(t)$$

General Tail Model of Poisson

In this model, arrivals and departures are combined based on Poisson's assumption, that is, times between arrivals and service times follow the exponential distribution. The development of the model is based on the long-term or steady-state behavior of the queue situation, achieved after the system has been in operation for a sufficiently long time. This type of analysis contrasts with the transient (or heating) behavior that prevails during the start of the system operation.

The general model assumes that both entry and exit rates depend on the state; which means that they depend on the number of customers in the service installation. For example, in a toll booth on a highway, managers tend to accelerate the collection of fees during peak hours. Define:

n = number of customers in the system (queuing, in addition to those being served).

Λ*n* = arrivals rate, if n customers are the system.

M*n* = Departure rate, if n customers are in the system.

Pn = stable status probability that n customers are in the system.

The general model derives Pn as a function of λn and μn. These probabilities are then used to determine the performance measures of the system, such as the average queue length, the average waiting time, and the average utilization of the installation. The probabilities Pn are determined by means of the transition rate diagram. The queue system is in state n when the number of clients in the system is n. for n> 0, the state n can change only to two possible states: n - 1 when an output occurs at the rate of μn, and n + 1 when an arrival occurs at the rate of λn. State 0 can only change to state 1 when an arrival occurs at the rate of λ0. Note that μ0 is undefined because no outputs can occur if the system is empty. Under steady state conditions, for n> 0, the expected rates of inflows to and from state n must be equal. Based on the fact that state n can change only to states n - 1 and n + 1, we have: (Expected input flow rate to the state) = λn - 1Pn - 1 + μn + 1Pn + 1 Also, (Expected output flow rate of state n) = (λn + μn) Pn By matching the two rates, we get the following balancing equation Eq. (11).

$$\lambda n-1Pn-1+\mu n+1Pn+1=(\lambda n\ +\ \mu n)Pn, n=1,2\ldots \qquad (11)$$

The balancing equation associated with n = 0 is

$$\lambda 0P0=\ \mu 1P1$$

The balancing equations are solved recursively as a function of P0. For n = 0, we have: Eq. (12)

$$P1=\left(\frac{\lambda 0}{\mu 1}\right)P0 \qquad (12)$$

Then, for n = 1, we have: Eq. (13)

$$\lambda 0P0+\mu 2P2=(\lambda 1+\mu 1)P1 \qquad (13)$$

Then, for n = 1, we have Substituting P1 = (λ0 / μ0) P0 and simplifying, we obtain Eq. (14):

$$P2=\left(\frac{\lambda 1\lambda 0}{\mu 2\mu 1}\right)P0 \qquad (14)$$

We can demonstrate by induction that Eq. (15)

$$Pn=\left(\frac{\lambda n-1\lambda n-2\ldots\lambda 0}{\mu n\mu n-1\ldots\mu 1}\right)P0, n=1,2,\ldots \qquad (15)$$

The value of P0 is determined with the equation Eq. (16)

$$\sum_{n=0}^{\infty}Pn=1 \qquad (16)$$

Specialized Poisson Tails

In the situation of specialized Poisson queues with c parallel servers. A client is selected from the queue to start the service with the first available server. The rate of arrivals to the system is λ clients per time unit. All parallel servers are identical, that is, the service rate of any server is μ clients per unit of time. The number of customers in the system is defined to include those in the service and those in the queue. A convenient notation to summarize the characteristics of the queue situation is given by the following format:

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

Where:

a = arrivals distribution.

b = distribution of outputs (service time).

c = number of parallel servers.

d = line discipline.

e = maximum number (finite or infinite) allowed in the system (queuing or in service).

f = Requesting font size (finite or infinite).

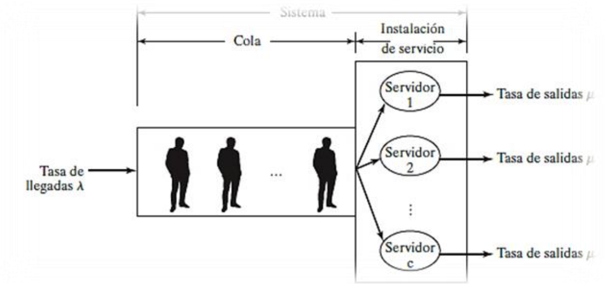

Figure 1 Schematic representation of a queue system with c parallel servers

As can be seen in figure two, it is a clear example of how a multi-server system works.

The standard notation to represent the distributions of arrivals and departures (symbols a and b) is:

M = Markovian (or Poisson) distribution of arrivals and departures (or equivalently exponential distribution of time between arrivals and service).

D = Constant time (deterministic).

Ek = Erlang distribution or time range (or equivalently, the sum of independent exponential distributions).

GI= General (generic) distribution of time between arrivals.

G = General (generic) distribution of service time.

The notation for queuing discipline (symbol d) includes:

FCFS = First to arrive, first to be served.

LCFS = Last to arrive, first to be served.

SIRO = Service in random order.

GD = General discipline (ie, any type of discipline).

To illustrate the use of the notation, the model (M / D / 10): (GD / 20 / q) uses Poisson arrivals (or time between exponential arrivals), constant service time, and 10 parallel servers. The discipline in queues is GD, and there is a limit of 20 clients throughout the system.

The font size from which customers arrive is infinite.

As a historical note, the first three elements of the notation (a / b / c) were devised by D.G. Kendall in 1953, and they are known in the literature as Kendall's notation. In 1966, A.M. Lee added the symbols d and e to the notation. This author added the last element, the symbol f, in 1968.

Measures of steady state performance

The most commonly used performance measures in a queue situation are:

Ls = Expected number of customers in a system.

Lq = Expected number of customers in a queue.

Ws = Timeout in the system.

Wq = Early waiting time in the queue.

Ĉ = Expected number of busy servers.

Remember that the system includes both the queue and the service facilities. We now demonstrate how these measurements are derived (directly or indirectly) from the probability of a stable state of n in the pn system as Eq. (17)

$$Ls = \sum_{n-1}^{\infty} npn \tag{17}$$

Eq. (18)

$$Lq = \sum_{n=c+1}^{\infty} (n-c)Pn \tag{18}$$

The relationship between Ls and Ws (also between Lq and Wq) is known as Little's formula and is given as Eq. (19):

$$Ls = \lambda efecW_s \tag{19}$$

Eq. (20)

$$Lq = \lambda efecW_q \tag{20}$$

These relationships are valid under rather general conditions. The λefec parameter is the effective arrival rate to the system. It is equal to the arrival rate λ (nominal) when all arriving customers can join the system. Otherwise, if some clients cannot join because the system is full (for example a parking lot), then λefec <λ. Later we will demonstrate how λefec is determined. There is also a direct relationship between Ws and Wq. By definition (Early waiting time in the system) = (early waiting time in the queue) + (operating time operated).

This translates as Eq. (21):

$$Ws= Wq + 1/\mu \qquad (21)$$

Then we can relate NO to La by multiplying both sides of the last formula by λefec, which together with the Little da formula: Eq. (22)

$$Ls = Lq + \lambda efec/ \mu \qquad (22)$$

The difference between the average amount in the system, Ls, and the average amount in the queue, Lq must be equal to the average number of servers occupied. Thus Eq. (23)

$$\hat{c} = Ls - Lq = \frac{\lambda efec}{\mu} \qquad (23)$$

Se deduce que (Installation use) = ĉ/c.

Single Server Models

Two models are presented for the case of a single server (c = 1). The first model does not limit the maximum number in the system, and the second represents a finite system limit. Both models assume an infinite capacity of the source. Arrivals occur at the rate of λ clients per unit of time and the service rate is μ clients per unit of time. **(M/M/1): (GD/q/q).** Using the general model notation, we have

$$\left.\begin{matrix}\lambda n=\lambda\\ \mu n=\mu\end{matrix}\right\}, n = 0,1,2,\dots$$

Even, λefec = λ and λ lost = 0, because all clients can join the system.

If p = λ / μ, the expression for Pn in the generalized model is reduced to

$$Pn = p^n P0, n = 0,1,2,\dots$$

To determine the value of p0 we use the identity Eq. (24)

$$p0(1 + p + p2 + \cdots) = 1 \qquad (24)$$

The sum of the geometric series is (1 / 1-p), provided that p <1. Therefore $P0 = 1 - p, p < 1$ Consequently, the following geometric distribution gives the general formula for pn Eq. (25)

$$Pn = (1 - p)p^n, n = 1,2,\dots(p < 1) \qquad (25)$$

The mathematical derivation of pn imposes the condition p <1, or λ <μ. If λ ≥ λ, the geometric series diverges, and the steady state probabilities pn do not exist. This result makes intuitive sense, because unless the service rate is greater than the arrival rate, the length of the queue will continue to grow and no stable state can be reached. The performance measure Lq is derived as follows Eq. (26):

$$Ls = \sum_{n=0}^{\infty} nPn = \sum_{n=0}^{\infty} n\,(1 - p)p^n \qquad (26)$$

$$= (1 - p)p\frac{d}{dp}\sum_{n=0}^{\infty} p^n$$

$$= (1 - p)p\frac{d}{dp}\left(\frac{1}{1-p}\right) = \frac{p}{1-p}$$

(M/M/1): (GD/N/∞). This model differs from (M / M / 1): (GD / q / q) in that there is a limit N on the number in the system (maximum queue length = N - 1). Some examples include manufacturing situations where a machine can have a limited intermediate space and a service window in your car in a fast food restaurant. New arrivals are not allowed when the number of customers in the system reaches N. Therefore,

$$\lambda n = \begin{cases} \lambda, & n = 0,1,\dots,N-1 \\ 0, & n = N, N+1 \end{cases}$$

$$\mu n = \mu,\ n = 0,1,\dots$$

Using p = λ / μ, the generalized model of the section gives:

$$n = \begin{cases} p^n p0, & n \le N \\ 0, & n > N \end{cases}$$

The value of p0 is determined from the equation $\sum_{n=0}^{\infty} Pn$ = 1, which gives Eq. (27)

$$P0(1 + p + p^2 + \cdots + p^N) = 1 \qquad (27)$$

O Eq. (28)

$$Pn\begin{cases}\frac{(1-p)p^n}{1-p^{N+1}}, p \neq 1\\ \frac{(1-p)p^n}{1-p^{N+1}}, p = 1\end{cases}\}, n = 0,1,\dots,N \qquad (28)$$

The value of p = λ / μ does not have to be less than 1 in this model, because the limit N controls the arrivals to the system. This means that λefec is the rate that matters in this case. Because customers get lost when there is N in the system, then,

$$\lambda perdida = \lambda pn$$
$$\lambda efec = \lambda - \lambda perdida = \lambda(1 - PN)$$

In this case, λefec <μ.

The expected number of customers in the system is calculated as: Eq. (29)

$$Ls = \sum_{n=0}^{N} nPn \qquad (29)$$

$$= \frac{1-p}{1-p^{N+1}} \sum_{n=0}^{N} np^n$$
$$= \left(\frac{1-p}{1-p^{N+1}}\right) p \frac{d}{dp} \sum_{n=0}^{N} p^n$$
$$= \frac{(1-p)p}{1-p^{N+1}} \frac{d}{dp} \left(\frac{1-p^{N+1}}{1-p}\right)$$
$$= \frac{p[1-(N+1)P^N + Np^{N+1}]}{(1-p)(1-p^{N+1})}, p \neq 1$$

Multi-server models

Three queue models with several parallel servers are considered. The first two models are the versions of several servers. The third model deals with the case of self-service, which is equivalent to having an infinite number of parallel servers. (M/M/c):(GD/∞/∞). This model deals with c identical parallel servers. The arrival rate is λ and the service rate per server is μ. In this situation λefec = λ because there is no limit on the number present in the system.

The effect of using c parallel identical servers is a proportional increase in the service rate of the installation. In terms of the generalized model, λn and μn are therefore defined as:

$$\lambda n = \lambda,\ n \geq 0$$
$$\mu n = \begin{cases} n\mu, & n < c \\ c\mu, & n \geq c \end{cases}$$

So, Eq. (30)

$$Pn = \begin{cases} \frac{\lambda^n}{\mu(2\mu)(3\mu)\dots(n\mu)} P0 = \frac{\lambda^n}{n!\mu^n} Po = \frac{p^n}{n!} P0, & n < c \\ \frac{\lambda^n}{(\Pi_{i=1}^{c} i\mu)(c\mu)^{n-c}} P0 = \frac{\lambda^n}{c!c^{n-c}\mu^n} Po = \frac{p^n}{c!c^{n-c}} P0, & n \geq c \end{cases} \qquad (30)$$

If p = λ / μ, and assuming that p / c <1, the value of p0 is determined from, $\sum_{n=0}^{\infty} pn = 1$ which gives, Eq. (31)

$$P0 \left\{\sum_{n=0}^{c-1} \frac{p^n}{n!} + \frac{p^c}{c!} \sum_{n=0}^{\infty} \left(\frac{p}{c}\right) n - c\right\} - 1$$
$$= \left\{\sum_{n=0}^{c-1} \frac{p^n}{n!} + \frac{p^c}{c!} \left(\frac{1}{1-\frac{p}{c}}\right)\right\} - 1, \frac{p}{c} < 1 \qquad (31)$$

The expression for Lq is determined as follows Eq. (32):

$$Lq = \sum_{n=c}^{\infty} (n-c)\, Pn \qquad (32)$$

$$= \sum_{k=0}^{\infty} kPk + c$$
$$= \sum_{k=0}^{\infty} k \frac{p^{k+c}}{c^k c!} P0$$
$$= \frac{p^{c+1}}{c!c} P0 \sum_{K=0}^{\infty} K \left(\frac{p}{c}\right) k - 1$$
$$= \frac{p^{c+1}}{c!c} P0 \frac{d}{d\left(\frac{p}{c}\right)} \sum_{K=0}^{\infty} \left(\frac{p}{c}\right) k$$
$$= \frac{p^{c+1}}{(c-1)!(c-p)^2} Po$$

Because λefec = λ, Ls = Lq + p. The measures Ws and Wq are determined by dividing Ls and Lq by λ.

Methodology

In order to meet the main objective of the project, it was first described the processes that are carried out related to container transfers and the customs control area. Below is a table representing the number of transfers made from January to December for four years.

Year	July	Aug	Sep	Oct	Nov	Dec
2018	115	90	171	142	224	248
2017	45	88	47	29	51	52
2016	137	95	25	31	181	90
2015	95	70	85	110	102	97
Total	392	343	328	312	558	487
Year	**July**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
2018	239	224	247	578	0	0
2017	46	90	156	112	119	124
2016	105	78	40	91	69	64
2015	113	139	229	247	254	127
Total	503	531	672	1028	442	315

Table 1 List of transfers made during 4 years

On the other hand, it was also necessary to observe the distribution of the containers in the yard, or, in the warehouse in case they required a service requested by the customer. They are stowed as follows:

- Three tall containers.
- They accommodate for days.

When they are downloaded to the warehouse they are distributed as follows:

- 13 containers per ship.
- They are placed from ship one to three for any service to perform.

The transfer time per container is 40 minutes to 1 hour depending on the position in which the container is located in the terminal from which it will be sent. A tract can transfer two 20-foot containers on its flat. A unit lasts in the module five minutes, which is what it takes to register:

- Check container seal against article 15.
- It is linked to the reference.
- Income and notification of sicrefis.

Based on the research carried out, it was found that the main problem is originated in the transfer of transfers normally for the entry of a unit takes half an hour to forty minutes but lately it took more than an hour to transfer a container and an increase of Containers which generated more time and more costs to customers so it was decided to implement the model of a single server and several servers to observe the level of service and observe the profitability of the company. The single server model will be used. It was applied to the following times.

Single server model

Model applied at forty minutes

We will convert the minutes to hours.

$$X = \frac{5\,minutes * 1\,hour}{60\,minutes} = \frac{5\,hours}{60} = 0.08\,hours.$$

$$X = \frac{40\,minutes * 1\,hour}{60\,minutes} = \frac{40\,hours}{60} = 0.66\,hours.$$

$$\lambda = \frac{1}{0.66\,h} = 1.5 \approx 2\,units.$$

$$\mu = \frac{1}{5\,min.} = 0.20\frac{units}{min}\left(\frac{60\,min}{hour}\right) = 12\,units\,per\,hour.$$

A) Average number of units in the system.

$$Ls = \frac{\lambda}{\mu-\lambda} = \frac{2}{12-2} = \frac{2}{10} = 0.2 \approx 0\,units.$$

B) Total time consumed by a unit in the module.

$$Ws = \frac{1}{\mu-\lambda} = \frac{1}{12-2} = \frac{1}{10} = 0.1\,hour.$$

C) System Usage Factor.

$$p = \frac{\lambda}{\mu} = \frac{2}{12} = 0.166 * 100 = 16.6\%$$

D) Average number of units queuing.

$$Lq = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{2^2}{12(12-2)} = \frac{4}{120} = 0.03 \approx 0\,units.$$

E) Probability that the module is empty.

$$po = 1 - p = 1 - 0.166 = 0.833 * 100 = 83.3\%$$

F) Probability that two units are found in the system.

$$P2 = (1 - 0.1666)(0.1666)2 = 0.023 * 100 = 2.31\%$$

G) Time in which customers wait in line.

$$Wq = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{2}{12(12-2)} = \frac{2}{120} = 0.01666666 * 60 = 1\,minute.$$

Model applied in half an hour

We will convert the minutes to hours

$$X = \frac{5\,minutes * 1\,hour}{60\,minutes} = \frac{5\,hours}{60} = 0.08\,hours.$$

$$X = \frac{30\,minutes * 1\,hour}{60\,minutes} = \frac{30\,hours}{60} = 0.50\,hours.$$

$$\lambda = \frac{1}{0.50\,h} = 1.5 \approx 2\,units.$$

$$\mu = \frac{1}{0.08\,h} = 13\,units\,per\,hour.$$

A) Average number of units in the system

$$Ls = \frac{\lambda}{\mu-\lambda} = \frac{2}{13-2} = \frac{2}{11} = 0.18 \approx 0\,units.$$

B) time consumed by a unit in the module.

$$Ws=\frac{1}{\mu-\lambda}=\frac{1}{13-2}=\frac{1}{11}=0.09\ hour.$$

C) System Usage Factor.

$$p=\frac{\lambda}{\mu}=\frac{2}{13}=0.1538*100=15.38\%$$

D) Average number of units queuing.

$$Lq=\frac{\lambda^2}{\mu(\mu-\lambda)}=\frac{2^2}{13(13-2)}=\frac{4}{143}=0.027\approx 0\ units.$$

E) E) Probability that the module is empty.

$$po=1-p=1-0.1538=0.8462*100=84.62\ \%$$

F) Probability that two units are found in the system.

$$P2=(1-0.1538)(0.1538)2=0.0199*100=1.99\ \%$$

G) Time in which customers wait in line.

$$Wq=\frac{\lambda}{\mu(\mu-\lambda)}=\frac{2}{13(13-2)}=\frac{2}{143}=0.0139*60=0.83\ hour\ \approx 1\ minute.$$

Multi-server model

With two servers

a) Probability that no unit is in the system.

$$P0=\frac{1}{\sum_{n=0}^{S-1}\frac{\left(\frac{\lambda}{\mu}\right)n}{n!}+\frac{\left(\frac{\lambda}{\mu}\right)s}{s!}\left[\frac{1}{1-\left(\frac{\lambda}{s\mu}\right)}\right]}$$

$$P0=\frac{1}{\sum_{n=0}^{1}\frac{\left(\frac{4}{12}\right)\wedge n}{n!}+\frac{\left(\frac{4}{12}\right)\wedge 2}{2!}\left[\frac{1}{1-\left(\frac{4}{2*12}\right)}\right]}$$

$$P0=\frac{1}{\frac{\left(\frac{4}{12}\right)\wedge 0}{0!}+\frac{\left(\frac{4}{12}\right)\wedge 1}{1!}+\frac{\left(\frac{4}{12}\right)\wedge 2}{2!}\left[\frac{1}{1-\frac{4}{24}}\right]}$$

$$P0=\frac{1}{1+0.3333+0.0555\left(\frac{1}{1-0.1666}\right)}=0.7248$$

b) Average number of units in the system.

$$Ls=\frac{\lambda\mu\left(\frac{\lambda}{\mu}\right)\wedge sP0}{(s-1)!(s\mu-\lambda)\wedge 2}+\frac{\lambda}{\mu}$$

$$Ls=\frac{4(12)\left(\frac{4}{12}\right)\wedge 2(0.7248)}{(2-1)!(2(12)-4)\wedge 2}+\frac{4}{12}$$

$$Ls=\frac{48(0.3333)\wedge 2(0.7248)}{400}+0.3333=0.3429\ units.$$

c) Average time in which a unit is within the system.

$$ws=\frac{0.3429}{4}=0.0857\ hours.$$

d) Number of units in the row.

$$Lq=P0\left[\frac{\left(\frac{\lambda}{\mu}\right)\wedge s+1}{(s-1)!\left(s-\frac{\lambda}{\mu}\right)\wedge 2}\right]$$

$$Lq=0.7248\left[\frac{0.3333^3}{(1)(2-0.3333)\wedge 2}\right]$$

$$Lq=0.7248\left[\frac{0.0370}{2.7778}\right]=0.1449\ units.$$

e) Waiting time in line.

$$Wq=ws-\frac{1}{\mu}=0.0857-\frac{1}{12}=0.002\ hours.$$

Model with four servers

a) Probability that no unit is in the system.

$$P0=\frac{1}{\sum_{n=0}^{S-1}\frac{\left(\frac{\lambda}{\mu}\right)n}{n!}+\frac{\left(\frac{\lambda}{\mu}\right)s}{s!}\left[\frac{1}{1-\left(\frac{\lambda}{s\mu}\right)}\right]}$$

$$P0=\frac{1}{\sum_{n=0}^{3}\frac{\left(\frac{5}{12}\right)\wedge n}{n!}+\frac{\left(\frac{5}{12}\right)\wedge 4}{4!}\left[\frac{1}{1-\left(\frac{5}{4*12}\right)}\right]}$$

$$P0=\frac{1}{\frac{\left(\frac{5}{12}\right)\wedge 0}{0!}+\frac{\left(\frac{5}{12}\right)\wedge 1}{1!}+\frac{\left(\frac{5}{12}\right)\wedge 2}{2!}+\frac{\left(\frac{5}{12}\right)\wedge 3}{3!}+\frac{\left(\frac{5}{12}\right)\wedge 4}{4!}\left[\frac{1}{1-\frac{5}{48}}\right]}$$

$$P0=\frac{1}{1+0.4166+0.0868+0.0120+0.0012\left(\frac{1}{1-0.1041}\right)}=0.6594$$

b) Average number of units in the system.

$$Ls=\frac{\lambda\mu\left(\frac{\lambda}{\mu}\right)\wedge sP0}{(s-1)!(s\mu-\lambda)\wedge 2}+\frac{\lambda}{\mu}$$

$$Ls=\frac{5(12)\left(\frac{5}{12}\right)\wedge 4\ (0.6594)}{(4-1)!(4(12)-5)\wedge 2}+\frac{5}{12}$$

$$Ls=\frac{60(0.4166)\wedge 4(0.6594)}{11{,}094}+0.4166=$$

$0.4167\ units.$

c) Average time in which a unit is within the system.

$$ws=\frac{0.4167}{5}=0.0833\ hours.$$

d) Number of units in the row.

$$Lq=P0\left[\frac{\left(\frac{\lambda}{\mu}\right)\wedge s+1}{(s-1)!\left(s-\frac{\lambda}{\mu}\right)\wedge 2}\right]$$

$$Lq = 0.6594\left[\frac{0.4166^5}{(6)(4-0.4166)\wedge 2}\right]$$

$$Lq = 0.6594\left[\frac{0.0125}{77.0412}\right] = 0.0001\,units.$$

e) Waiting time in line.

$$Wq = ws - \frac{1}{\mu} = 0.0833 - \frac{1}{12} = 0\ hours.$$

Results

When carrying out the transfer of the containers by transfer, different times were usually carried out, but lately they were delayed to 1 hour 20 minutes which generated more costs since it caused the delay in the operations. Therefore, the model of a single server was applied at different times to analyze the behavior of said transfers and observe how long the unit is registered in order to enter the warehouse or yard where the download is carried out according to the service programmed by the client. The multi-server model was also developed.

Model applied at forty minutes

Landa throws us that it is possible to enter two units per hour and mu indicates the speed at which the server can serve units which threw us 12 units per hour, the average number of units in the system throws us 0 units, the time that provides the module to serve a unit is six minutes per unit, there is 16.6% of the system being in use when a unit arrives and 0 units would be queuing in the system since the unit present in module will be attended and the possibility arises that the system is empty with 83.3% and gives us a 2.31% probability that two units are found in the system.

Model applied in half an hour

Landa tells us that two units can be served per hour and mu represents that 13 units can be served per hour and there are 0 units in the system and the time elapsed by registering a unit is 5 minutes and there is a 15.83 % probability that the system is in use and that no unit is found in the system lining up and there is 84.62% that the input module is empty when a unit is being entered and with a 1.62% probability that two units are found in the system.

With this resolution in methods there is not a big difference in the time of forty minutes and half an hour, so it will be possible to enter between that time the number of transfers during an eight hour shift, which reduces the costs to customers by operators, machinery and units used in overtime. If the arrival speed of units were greater than the service speed, what would cause the queue to grow infinitely and the system would become saturated and cause the service to be delayed and more costs will be generated due to delayed operation. The multi-server model was also applied to analyze the level of service since the organization only has one server. The multi-server model was applied to two servers and to four servers performing the analysis, the proposal is given:

- Have two servers to meet the demand for service.

- Acquire two more units for the transfer of containers by transfer or loose cargo.

- Hire the services of a carrier when there is a high number of transfers.

By acquiring two more units you would already have four units and over time you can save the costs of contracting the services of a third party to cover the demand for transfers and you can also provide the service of moving empty containers from terminal to terminal, the number of customers will be increased and revenues would increase to 40% over a period of two years.

By having two servers, you avoid generating a long line of units waiting for their entry either transfer, loose cargo, vehicles, container dispatch, the more this generates a bottleneck and therefore the operation is delayed.

Conclusion

After analyzing the possible models to be developed and carrying out the application of the mathematical models of the theory of selected queues that for this case was of one server and several servers, it was possible to analyze the time in which it is carried out the registration of a unit in modules, the waiting times, the possibilities that the system is empty, that a bottleneck may arise.

Also, the possibilities of increasing the number of customers and improving the level of service provided and keeping a better control in all the developed processes were found, allowing a better fluidity in the input and output modules, as well as in the operational area. By having two servers, the waiting time in the units is reduced by 30%, there is even a 70% chance that when a unit is present, no unit is found on the server and there are not many delays in the service provided to the clients and it is finished in a timely manner and it is possible not to generate more time and more costs to the client and the bottlenecks are eliminated.

References

1. *EAE business school.* (29 de Noviembre de 2017). Obtenido de https://retos-operaciones-logistica.eae.es/todo-lo-que-no-sabias-de-la-carta-de-porte/

2. edición., t. H. (2004). *scribd.* Obtenido de https://es.scribd.com/doc/112051798/Toma-de-Decisiones-Bajo-Certidumbre-Riesgo-e-Incertidumbre

3. Gomez, R. A. (19 de Noviembre de 2010). *metodos cuantitativos utilizados en diseños de la gestión de almacenes y centros de distribución.*

4. ley aduanera . (ultima reforma publicada DOF 25-06-2018). En *ley aduanera* .

5. Matias, M. F. (23 de Mayo de 2005). *gestiopolis* . Obtenido de https://www.gestiopolis.com/teoria-de-colas/

6. Moody, P. E. (1990). *toma de decisiones gerenciales* . COLOMBIA: MCGRAW HILL.

7. peña, o. n. (s.f.). *optimización de la gestion de inventarios en la sucursal* .

8. RAMIREZ, A. C. (2009). *MANUAL DE GESTIÓN LOGISTICA DEL TRANSPORTE Y DISTRIBUCIÓN DE MERCANCIA.* COLOMBIA : UNINORTE.

9. *REGLAS GENERALES DE COMERCIO EXTERIOR.* (27 de ENERO de 2016). Obtenido de DIARIO OFICIAL.

10. sabater, j. p. (2015). *aplicando teoria de colas en dirección de operaciones* .

11. Taha, H. A. (2004). *INVESTIGACIÓN DE OPERACIONES 7A. EDICIÓN.* . México: Pearson .

12. Taha, H. A. (2012). *investigación de operaciones. Novena edición.* . México: Pearson Educacion .
TIBA MEXICO . (17 de ENERO de 2015). Obtenido de https://www.tibagroup.com/mx/glosario-de-terminos-maritimos-portuarios