

# OPTIMIZACIÓN DE SISTEMAS DE COLAS CON SERVICIOS MASIVOS DE CAPACIDAD FIJA

MIGUEL MIRANDA

Facultad de Ingeniería. Universidad de Buenos Aires. ARGENTINA<sup>1</sup>  
*miguelmiranda@netscape.net*

Fecha recepción: Marzo 2016 - Fecha aprobación: Noviembre 2016

## RESUMEN

Este trabajo aborda un método para determinar cuantitativamente, entre otras variables, el número promedio de clientes en un sistema de procesamiento en masa, en donde los usuarios arriban individualmente a intervalos aleatorios y son atendidos en grupos de tamaño fijo a una tasa de servicio Poisson. Para los sistemas considerados, de un solo canal y sin restricciones de lugar de espera, se define un factor de demora que permite determinar el número promedio de clientes que están esperando a que el canal comience el servicio. Con estas variables se puede formular una expresión de la longitud promedio del sistema y optimizar el tamaño del lote de procesamiento, en términos de costos de espera, de estructura, de tamaño y de producción.

**Palabras Claves:** sistemas masivos de colas P/P/1 – servicios en masa - optimización del lote.

## ABSTRACT

This paper presents a method to quantitatively determine, among other variables, the average number of customers in a bulk service queue system, wherein customers arrive individually at random time intervals and are simultaneously served in groups of fixed size at a Poissonian rate. Consideration is given to single-server systems with unlimited buffer, wherein a delay factor, which allows to determine the average number of clients waiting for the channel to start processing, is defined. With these two variables it is possible to formulate a quantitative expression for the system average length and thus allowing to optimize the processing bulk size, in terms of waiting, operating, sizing and producing costs.

**Key Words:** M/M/1 bulk queueing systems - bulk service queues - batch optimization.

## 1. INTRODUCCIÓN

Las instalaciones en las cuales los clientes son atendidos en lotes se denominan sistemas de servicios grupales o masivos. El tamaño del lote de

---

<sup>1</sup> También Facultad de Ingeniería y Ciencias Agrarias – Universidad Católica Argentina – Argentina.

procesamiento puede ser fijo, generalmente determinado por la capacidad de clientes que los canales pueden atender simultáneamente, o variable entre un número mínimo y otro máximo. A diferencia de los procesos de nacimiento muerte, los sistemas de eventos masivos, presentan una considerable dificultad para un enfoque analítico, especialmente los que admiten infinitos estados posibles, por lo que típicamente se recurre a la técnica de la simulación como herramienta de solución.

En este trabajo se abordarán modelos cuantitativos para formular sistemas en donde los clientes arriban individualmente, a intervalos de tiempo aleatorios conforme a un proceso Poisson, forman una cola simple no restringida, y se atienden en lotes de igual tamaño en un canal multiprocesador único, siendo la duración del servicio prestado una variable aleatoria exponencialmente distribuida. A diferencia de los métodos conocidos se propone aquí un procedimiento para determinar expresiones de las variables características del sistema de una forma simple, directa e independiente de la determinación previa de la distribución de probabilidades de estado.

Un objeto principal de este trabajo consiste en determinar los tamaños óptimos de los lotes a fin de minimizar el costo total esperado para un período de tiempo, que contemple la permanencia de clientes en el sistema, la estructura del canal, el procesamiento de cada lote y los recursos insumidos en el servicio que se presta.

A tal fin, el modelo matemático sugiere la determinación de un factor de eficiencia, proporcionándose expresiones analíticas y tablas para su cálculo, a partir del cual se puede establecer el tiempo de permanencia promedio en el sistema y el resto de las variables de eficiencia.

La nomenclatura utilizada en este trabajo es la habitual de Teoría de Colas (Miranda, 2013). En particular, se utilizarán indistintamente los vocablos “cliente” y “usuario” para nombrar a las entidades (personas, productos, etc.) que arriban al sistema con el fin de recibir un servicio o proceso en la instalación; y las palabras “canal” o “servidor” para denominar al recurso (equipo, máquina, etc.) que presta el servicio o produce el trabajo.

## 2. SERVICIOS MASIVOS DE TAMAÑO FIJO

Muchos sistemas de producción y de transporte proveen el servicio en forma conjunta a una cantidad exacta de clientes, como sería el caso de un equipo de tratamiento térmico con una capacidad para procesar una cantidad  $K$  de unidades que llegan al sector a medida que son fabricadas, y en donde, por razones de costo operativo, el equipo no se pone en funcionamiento mientras no se haya completado su capacidad. Si bien en la práctica hay diversidad de instalaciones con múltiples canales multiprocesadores, se investigará en este estudio el caso general de sistemas constituidos por un solo canal. En la notación Kendall<sup>2</sup>, el tamaño fijo se consigna en el superíndice del código dispuesto en la segunda posición, que corresponde al tipo de proceso del servicio.

### 2.1. Hipótesis

Consideraremos un sistema  $P/P^K/1$  en régimen permanente, con clientes que no están sujetos a impaciencia. Cuando el canal finaliza un servicio, si en cola hay un número igual o superior a una cantidad  $K$  de clientes, comienza inmediatamente el próximo servicio; en caso contrario, el canal permanecerá inactivo hasta tanto se haya conformado un grupo de clientes de tamaño  $K$ .

### 2.2. Parámetros

- $K$ : Tamaño del lote de servicio [clientes/lote]. Normalmente es igual a la capacidad del procesamiento del canal.
- $\lambda$ : Tasa promedio de arribo de clientes al sistema [clientes/tiempo]: Su inversa es el intervalo promedio entre arribos de clientes:  $T_A = 1/\lambda$ .
- $\mu$ : Tasa o velocidad promedio de procesamiento de lotes [clientes/lote]:  $T_S = 1/\mu$ . Su inversa es la duración promedio del servicio.
- $\mu_K$ : Velocidad promedio de procesamiento de clientes [clientes/tiempo]:  $\mu \cdot K$ .
- $\rho$ : Factor de tráfico [clientes/lote]: Relación entre la tasa de ingreso de clientes y la tasa de servicio de lotes:  $\rho = \lambda/\mu$ .
- $\rho_K$ : Factor de servicio [adimensional]: Cociente entre la tasa de ingreso de clientes al sistema y la velocidad de servicio de clientes del canal.

### 2.3. Variables

- $\psi$ : Factor de demora.

<sup>2</sup> Esta notación tiene 6 posiciones:  $x/y/M/N/z/(N')$ , en donde 'x' se refiere el tipo de proceso de arribos al sistema, por ejemplo P (Poisson),  $E_K$ (Erlang  $K$ ), D (Determinístico), Hx (hiperexponencial), Hi (hipoexponencial), G (general), etc.; 'y' se refiere el tipo de proceso de servicios en los canales; M indica la cantidad de canales en paralelo; N indica el número máximo de clientes que puede residir en el sistema (infinito por default); 'z' señala la modalidad de atención (FIFO por default); y N' indica el tamaño de la población (infinito, por default).

- $p(n)$ : Probabilidad de que en el sistema haya 'n' clientes.
- L: Longitud promedio del sistema.
- $L_C$ : Longitud promedio de cola.
- B: Número esperado de clientes dentro del canal mientras éste está inactivo, esperando que se forme un lote procesable.
- $P_A$ : Porcentaje de actividad del canal.
- H: Número promedio de clientes recibiendo el servicio en el sistema.
- $\bar{\lambda}$ : Tasa promedio de ingreso de clientes al sistema.
- $\bar{\mu}$ : Tasa o velocidad promedio de egreso de clientes del sistema.
- $W_C$ : Tiempo promedio de espera de un cliente en cola.
- W: Tiempo promedio de permanencia de un cliente en el sistema.

### 2.4. Representación gráfica

En el gráfico de la Figura 1 se muestra esquemáticamente un sistema  $P/P^3/1$ , cuyo canal está subdividido en los 3 lugares disponibles para alojar a los clientes que se deben procesar.

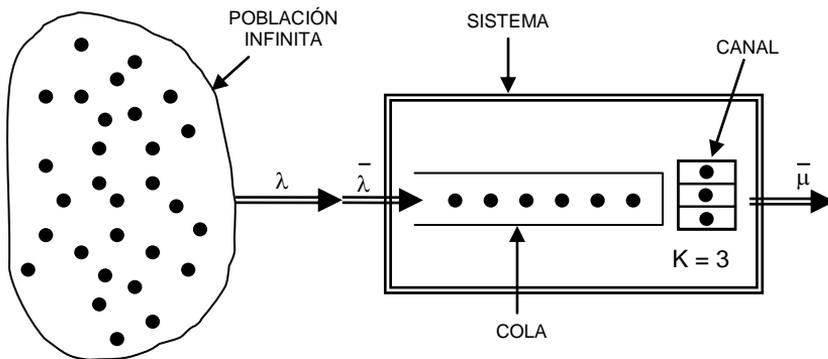


Figura 1: Sistema  $P/P^3/1$

### 2.5. Modelización

Siendo el intervalo entre arribos de clientes una variable con distribución exponencial, el tiempo requerido para que se conforme un grupo procesable (de K clientes) en el sistema tendrá una distribución Erlang-K, de media:

$$T_K = \sum_{1}^K T_A = K \cdot T_A = \frac{K}{\lambda}$$

cuya inversa es la cantidad promedio de lotes que se conforman por unidad de tiempo:

$$\lambda_K = \frac{\lambda}{K}$$

El factor de servicio está dado por cualquiera de las siguientes

relaciones:

$$\rho_K = \frac{\lambda}{\mu_K} = \frac{\lambda_K}{\mu} = \frac{\lambda}{K \cdot \mu} = \frac{\rho}{K} \quad [1]$$

El canal estará inactivo siempre que en el sistema haya una cantidad de clientes menor que el tamaño de un lote procesable. En consecuencia, el porcentaje de actividad será igual a la probabilidad de que en el sistema haya una cantidad mayor o igual a "K" clientes:

$$P_A = \sum_K^{\infty} p(n) = 1 - \sum_0^{K-1} p(n) \quad [2]$$

El número promedio de clientes recibiendo atención en el canal será igual al tamaño del lote de clientes multiplicado por el porcentaje de actividad del canal:

$$H = K \cdot P_A \quad [3]$$

Dado que se considera que no hay restricciones de capacidad de cola ni impaciencia, la tasa de ingresos es igual a la tasa de arribos:

$$\bar{\lambda} = \lambda \quad [4]$$

Por su parte, el número promedio de clientes atendidos por unidad de tiempo es igual a la velocidad promedio de atención de lotes multiplicada por el número promedio de clientes que están recibiendo atención. En consecuencia:

$$\bar{\mu} = \mu \cdot H = \mu \cdot K \cdot P_A \quad [5]$$

En régimen permanente, la tasa de ingreso de clientes debe ser igual a la tasa de egresos:  $\bar{\lambda} = \bar{\mu}$ . Entonces, combinando [4] y [5], y despejando:

$$\lambda = K \cdot \mu \cdot P_A$$

$$P_A = \frac{\rho}{K} = \rho_K \quad [6]$$

Es decir, el porcentaje de actividad del canal es igual al factor de servicio del sistema; y también, comparando [3] y [6], tenemos que el número promedio de clientes atendidos en el canal es igual a la relación de tráfico del canal  $\lambda/\mu$ :

$$H = \rho \quad [7]$$

Obviamente, para que el sistema se encuentre en equilibrio la velocidad de procesamiento de clientes debe ser mayor a la tasa de arribos, ya que de otra forma el sistema no tendría capacidad de atención suficiente para atender la demanda del servicio que se presta, y la cola crecería indefinidamente:

$$\lambda < \mu_K = K \cdot \mu \quad \rightarrow \quad \rho_K = \frac{\rho}{K} < 1 \quad \rightarrow \quad \rho = H < K \quad [8]$$

### 2.5.1. Factor de demora

El canal del sistema puede estar inactivo (desocupado o parcialmente

ocupado) o activo (completamente ocupado). El factor de servicio mide el nivel de ocupación activa del sistema, y constituye un índice de funcionalidad del sistema. Un valor alto de  $\rho_K$  puede ser deseable para la administración, pero no para los usuarios, quienes deben esperar en cola más tiempo. Más aún, como medida de eficiencia desde el punto de vista de los clientes, el factor de servicio es insuficiente, dado que no considera el tiempo adicional que deben esperar algunos de ellos mientras el canal esté inactivo a la espera de completar su capacidad. Por ejemplo, un sistema con un  $K$  alto genera alto tiempo ocioso del canal y atraso en la reanudación del servicio. Definiremos, entonces, un factor de demora, indicándolo con la letra griega  $\psi$ , que tenga en cuenta tanto el grado de ocupación activa del canal como el de ocupación inactiva. Este índice opera como un factor de servicio solamente cuando el estado del sistema es igual o mayor a  $K$ , por lo que la relación de probabilidades de estado de un estado  $n$  cualquiera y del estado  $n-1$  es:

$$\frac{p(n)}{p(n-1)} = \psi \quad \text{para } n \geq K \tag{9}$$

Dado que las probabilidades de la sucesión de los estados deben ser decrecientes a partir de  $n \geq K$  para que el sistema alcance el estado de equilibrio, se tiene que verificar la condición:  $\psi < 1$ . En caso contrario, el sistema no estaría bien dimensionado (en términos de velocidad de procesamiento o de tamaño de lote de procesamiento) para satisfacer la demanda del servicio que presta, y así la longitud de la cola crecería indefinidamente. Un valor  $\psi$  bajo (como consecuencia de una alta velocidad de servicio respecto a la tasa de llegadas, o de un  $K$  bajo, o de ambas condiciones) implica un sistema eficiente para el usuario.

Para determinar el valor del factor de demora se analizará la cadena de Markov (Bolch, 2006) correspondiente al sistema  $P/P^K/1$ , que se muestra en la Figura 2:

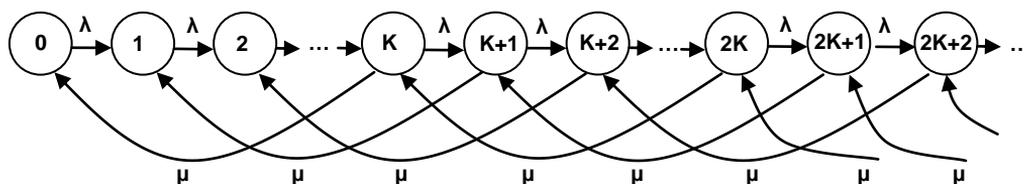


Figura 2: Cadena markoviana genérica de un sistema  $P/P^K/1$

Haciendo los balances probabilísticos en los nodos, tendremos las expresiones de las ecuaciones de estado:

$$p(0) \cdot \lambda = p(K) \cdot \mu \quad \text{para } n = 0 \tag{10}$$

$$p(n) \cdot \lambda = p(K+n) \cdot \mu + p(n-1) \cdot \lambda \quad \text{para } 0 < n < K \tag{11}$$

$$p(n) \cdot (\lambda + \mu) = p(K+n) \cdot \mu + p(n-1) \cdot \lambda \quad \text{para } n \geq K \tag{12}$$

las que, conjuntamente con la condición de colectividad exhaustiva:

$$\sum_{n=0}^{\infty} p(n) = 1$$

permitirían formular las funciones generativas para resolver el problema. Este método es complejo y engorroso, por lo que se procederá de otra forma más directa y sencilla. Para ello se determinará el factor de demora en función de los parámetros del problema.

Se indican a continuación las relaciones entre la probabilidad de un estado  $n > K$  y la probabilidad del estado  $n = K$ , en función del factor de demora:

$$\begin{aligned} \frac{p(n)}{p(n-1)} = \psi & \quad \rightarrow \quad \frac{p(K+1)}{p(K)} = \psi \\ \frac{p(K+2)}{p(K+1)} = \frac{p(K+2)}{p(K) \cdot \psi} = \psi & \quad \rightarrow \quad \frac{p(K+2)}{p(K)} = \psi^2 \\ \frac{p(K+3)}{p(K+2)} = \frac{p(K+3)}{p(K) \cdot \psi^2} = \psi & \quad \rightarrow \quad \frac{p(K+3)}{p(K)} = \psi^3 \end{aligned}$$

y así, por inducción completa, tendremos que:

$$\frac{p(K+3)}{p(K+2)} = \frac{p(K+3)}{p(K) \cdot \psi^2} = \psi \quad \rightarrow \quad \frac{p(K+n)}{p(K)} = \psi^n$$

Pero, teniendo en cuenta [10], la sucesión la podemos expresar también en función de la probabilidad de que el sistema esté vacío:

$$\frac{p(K+n)}{\rho \cdot p(0)} = \psi^n \quad \rightarrow \quad p(K+n) = \rho \cdot \psi^n \cdot p(0)$$

Es decir:

$$\boxed{p(n) = \rho \cdot \psi^{n-K} \cdot p(0)} \quad \text{para } n \geq K: \quad [13]$$

Del mismo modo, la relación entre la probabilidad de un estado  $p(n)$  y  $p(0)$  para  $0 < n < K$ , teniendo en cuenta las ecuaciones [10] y [11] es:

$$\frac{p(n)}{p(0)} = \frac{p(K+n) \cdot \mu + p(n-1) \cdot \lambda}{p(K) \cdot \mu} = \frac{p(K+n)}{p(K)} + \frac{p(n-1) \cdot \rho}{p(0) \cdot \rho} = \frac{p(K+n)}{p(K)} + \frac{p(n-1)}{p(0)}$$

Es decir:

$$\frac{p(n)}{p(0)} = \frac{p(K+n)}{p(K)} + \frac{p(n-1)}{p(0)} \quad [14]$$

Poniendo ahora esta relación en función del factor de demora para cada estado y despejando:

$$\frac{p(1)}{p(0)} = \frac{p(K+1)}{p(K)} + \frac{p(0)}{p(0)} = \psi + 1 \quad \rightarrow \quad p(1) = p(0) \cdot (1 + \psi)$$

$$\frac{p(2)}{p(0)} = \frac{p(K+2)}{p(K)} + \frac{p(1)}{p(0)} = \psi^2 + \psi + 1 \quad \rightarrow \quad p(2) = p(0) \cdot (1 + \psi + \psi^2)$$

$$\frac{p(3)}{p(0)} = \frac{p(K+3)}{p(K)} + \frac{p(2)}{p(0)} = \psi^3 + \psi^2 + \psi + 1 \quad \rightarrow \quad p(3) = p(0) \cdot (1 + \psi + \psi^2 + \psi^3)$$

Y así, en forma genérica:

$$p(n) = p(0) \cdot \sum_{k=0}^n \psi^k \quad \text{para } n < K \quad [15]$$

Teniendo en cuenta ahora [9] para  $n = K$ :

$$p(K) = \psi \cdot p(K-1) = \psi \cdot p(0) \cdot \sum_{k=0}^{K-1} \psi^k = \psi \cdot p(0) \cdot \left( \sum_{k=1}^{K-1} \psi^k + 1 \right) = \left( \sum_{k=1}^{K-1} \psi^{k+1} + \psi \right) \cdot p(0)$$

Luego:

$$p(K) = \sum_{k=1}^K \psi^k \cdot p(0)$$

y, por [10] se concluye que:

$$\rho = \sum_{k=1}^K \psi^k \quad [16]$$

Finalmente, desarrollando esta serie geométrica:

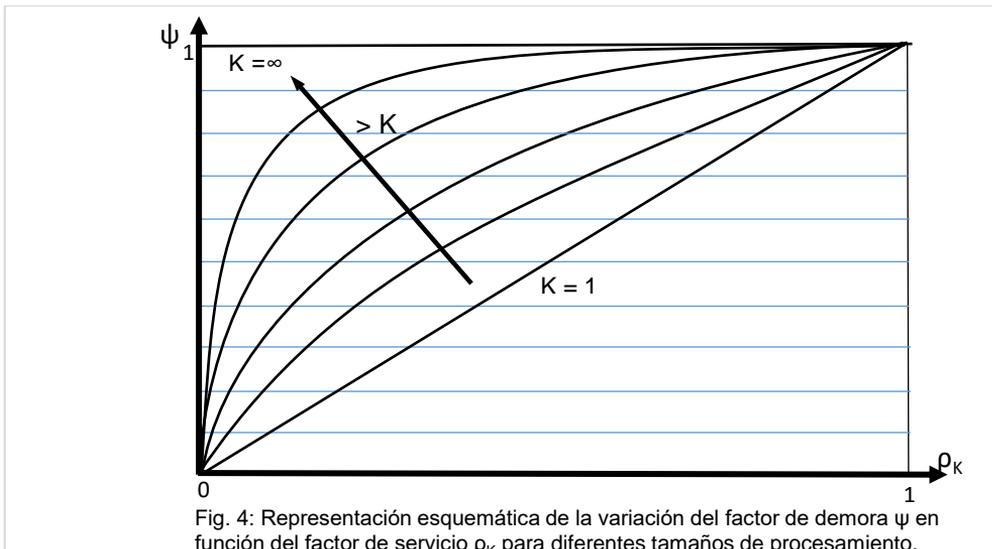
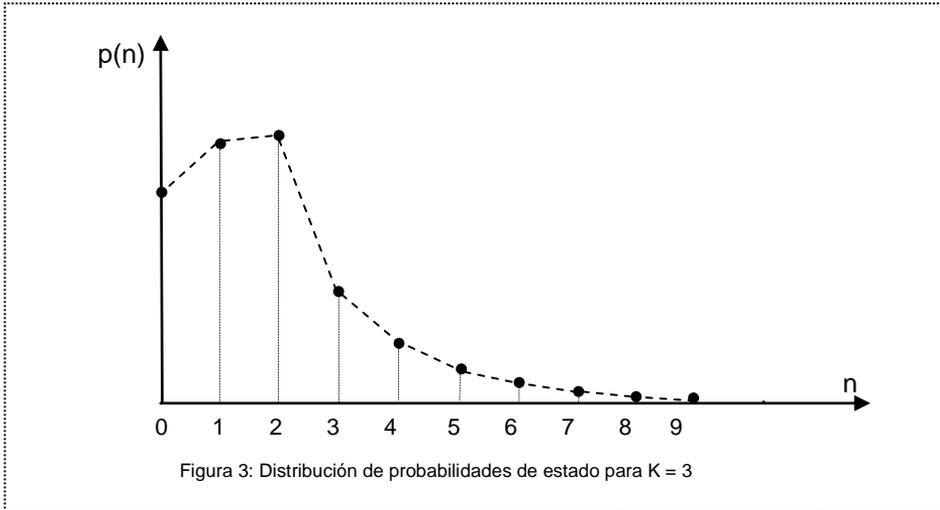
$$\rho = \frac{\psi \cdot (1 - \psi^K)}{1 - \psi} \quad [17]$$

Para obtener el factor de demora  $\psi$  se debe resolver la ecuación de grado  $K$  [17], que se puede hacer con cualquier sistema de programación matemática o utilizando la Tabla 1 en la que se proporcionan valores calculados de  $\psi$  para distintas combinaciones de  $\rho_K$  y  $K$ , hasta un tamaño de lote de 12 unidades, variando  $\rho_K$  entre 0 y 1, con incrementos de 0,05 puntos. La determinación de esta variable permitirá obtener analíticamente los valores de las variables características de manera muy sencilla, como se verá más adelante en la sección 2.5.4.

Observando la expresión [17] surge, obviamente, que para el caso de servicios individuales ( $K=1$ ) el factor de demora es igual al factor de servicio.

La Figura 3 muestra esquemáticamente la función de densidad de probabilidades de estado, para un sistema  $P/P^3/1$ . Esta función es creciente hasta el estado  $n = K-1$  y luego decreciente, tal como surge de las ecuaciones [15] y [14], respectivamente. La variación del factor de demora  $\psi$  en función del

porcentaje de utilización del canal (es decir, del factor de servicio de clientes  $\rho_K$ ) se muestra en la Figura 4 para diferentes tamaños de procesamiento.



### 2.5.2 Determinación de las probabilidades de estado en función de $\psi$

Para  $n \geq K$ , conforme [2] y [6]:

$$\sum_{n=K}^{\infty} p(n) = \frac{\rho}{K}$$

Además, teniendo en cuenta [13]:

$$p(0) \cdot \rho \cdot \sum_{n=K}^{\infty} \psi^{n-K} = \frac{\rho}{K}$$

La serie tiene convergencia en:

$$\sum_{n=K}^{\infty} \psi^{n-K} = \frac{1}{1-\psi}$$

Por lo tanto:

$$p(0) \cdot \rho \cdot \frac{1}{1-\psi} = \frac{\rho}{K}$$

Entonces la expresión de la probabilidad de que el sistema esté vacío, en función de  $\psi$  y  $K$  es:

$$\boxed{p(0) = \frac{1-\psi}{K}} \quad [18]$$

Teniendo en cuenta [18], al reemplazar los distintos valores de  $n < K$  en la ecuación [15], tendremos:

$$p(1) = \frac{1-\psi}{K} \cdot (1+\psi) = \frac{1-\psi^2}{K}$$

$$p(2) = \frac{1-\psi}{K} \cdot (1+\psi+\psi^2) = \frac{1-\psi^3}{K}$$

$$p(3) = \frac{1-\psi}{K} \cdot (1+\psi+\psi^2+\psi^3) = \frac{1-\psi^4}{K}$$

...

Así, genéricamente:

$$\boxed{p(n) = \frac{1-\psi^{n+1}}{K}} \quad \text{para } n < K \quad [19]$$

En el caso particular de que  $n = K-1$ :

$$p(K-1) = \frac{1-\psi^K}{K}$$

Teniendo en consideración ahora la ecuación [9], para  $n = K$ ,

$$\frac{p(K)}{p(K-1)} = \psi \quad \rightarrow \quad p(K) = \frac{1-\psi^K}{K} \cdot \psi$$

Para distintos valores de  $n$  superiores a  $K$  tendremos:

$$p(K+1) = \frac{1-\psi^K}{K} \cdot \psi^2 \quad \rightarrow \quad p(K+2) = \frac{1-\psi^K}{K} \cdot \psi^3$$

....

y, en general:

$$p(n) = \frac{1 - \psi^K}{K} \cdot \psi^{n-K+1} \quad \text{para } n \geq K \quad [20]$$

### 2.5.3 Determinación de la variable B en función de $\psi$

La cantidad promedio de clientes esperando dentro del canal hasta que se forme un lote procesable, mientras el sistema se encuentra sin funcionar, es igual a la esperanza matemática de que haya “n” clientes en el sistema cuando el canal está inactivo:

$$B = \sum_{n=0}^{K-1} n \cdot p(n) \quad [21]$$

Reemplazando [19] en [21]:

$$B = \sum_{n=0}^{K-1} n \cdot \frac{1 - \psi^{n+1}}{K} = \sum_{n=0}^{K-1} \frac{n - n \cdot \psi^{n+1}}{K}$$

Entonces:

$$B = \frac{\sum_{n=0}^{K-1} n - \sum_{n=0}^{K-1} n \cdot \psi^{n+1}}{K} \quad [22]$$

La expresión de la progresión aritmética que constituye el primer término del numerador es:

$$\sum_{n=0}^{K-1} n = (K-1) \cdot \frac{1 + (K-1)}{2} = (K-1) \cdot \frac{K}{2} \quad [23]$$

Por otra parte, para desarrollar la serie del segundo término:

$$\sum_{n=0}^{K-1} \psi^n = \frac{1 - \psi^K}{1 - \psi}$$

Derivando ambos miembros por “n” y ordenando:

$$\sum_{n=0}^{K-1} n \cdot \psi^{n-1} = \frac{(-K \cdot \psi^{K-1}) \cdot (1 - \psi) + (1 - \psi^K)}{(1 - \psi)^2} = \frac{1 + \psi^K \cdot (K-1) - K \cdot \psi^{K-1}}{(1 - \psi)^2}$$

Multiplicando ahora por  $\psi^2$ , también en ambos miembros, se arriba a la expresión de la serie buscada:

$$\sum_{n=0}^{K-1} n \cdot \psi^{n+1} = \frac{1 + \psi^K \cdot (K-1) - K \cdot \psi^{K-1}}{(1 - \psi)^2} \cdot \psi^2 \quad [24]$$

Reemplazando [23] y [24] en [22], tendremos la fórmula del número

promedio de clientes en espera a que el canal se complete con la cantidad requerida para iniciar el proceso, en función del factor de demora:

$$B = \frac{(K-1)}{2} - \frac{1+(K-1) \cdot \psi^K - K \cdot \psi^{K-1}}{(1-\psi)^2} \cdot \frac{\psi^2}{K} \quad [25]$$

2.5.4. Determinación y cálculo de las variables características

Considerando la ley de Little (Little, 1961), la propiedad PASTA<sup>3</sup> (Whit, 1990) y la característica de falta de memoria de los procesos Poisson (Gross, 1988), la expectativa de permanencia que tendrá un cliente recién arribado al sistema será igual a la suma de:

- el tiempo promedio que deberá esperar hasta que se complete un grupo procesable de K clientes.

$$W_B = \frac{B}{\lambda} = B \cdot T_A \quad [26]$$

- el tiempo promedio de espera en cola, es decir el tiempo que transcurre hasta que se atiendan los clientes que estaban delante de él en el sistema, excepto aquellos que se van a atender conjuntamente con el cliente<sup>4</sup>.

$$W_C = \left( L - \frac{K-1}{2} \right) \cdot \frac{T_S}{K} \quad [27]$$

- el tiempo promedio del servicio del grupo en el que se atiende el cliente:  $T_S$

Luego, integrando los tres tiempos, se tiene el tiempo promedio de permanencia:

$$W = W_B + W_C + T_S$$

$$W = B \cdot T_A + \left( L - \frac{K-1}{2} \right) \cdot \frac{T_S}{K} + T_S = B \cdot T_A + L \cdot \frac{T_S}{K} - \frac{K-1}{2} \cdot \frac{T_S}{K} + T_S \quad [28]$$

Aplicando Little a la longitud promedio del sistema:

<sup>3</sup> PASTA (*Poisson Arrivals See Time Averages*): Esta propiedad establece que los clientes que arriban a un sistema conforme a un proceso Poisson perciben al sistema como lo haría un observador externo del sistema; es decir, tienen en cuenta los valores promedios de los tiempos, por lo que la probabilidad de que el ingresante encuentre a su llegada que en el sistema hay n clientes es igual a la probabilidad de que el sistema se encuentre en el estado n. En virtud de esta propiedad, el número esperado de clientes que un recién llegado encuentra en el sistema es L.

<sup>4</sup> Cuando el cliente ingresante reciba la atención lo hará con “K-1” compañeros de grupo. Algunos de ellos habrán arribado antes que él y otros luego. El número esperado de compañeros que estaban delante de él en cola es: (K-1)/2.

$$W = B \cdot T_A + \lambda \cdot W \cdot \frac{T_S}{K} - \frac{K-1}{2} \cdot \frac{T_S}{K} + T_S$$

Operando:

$$W \cdot \left(1 - \frac{\rho}{K}\right) = B \cdot T_A - \frac{K-1}{2} \cdot \frac{T_S}{K} + T_S$$

$$W = \frac{B \cdot T_A - \frac{K-1}{2} \cdot \frac{T_S}{K} + T_S}{\left(1 - \frac{\rho}{K}\right)}$$

$$W = \left( B \cdot T_A - \frac{K-1}{2} \cdot \frac{T_S}{K} + T_S \right) \cdot \frac{K}{K-\rho} \quad [29]$$

Observar que para  $K = 1$  (es decir procesamiento individual) la expresión anterior queda igual a la del caso P/P/1:

$$W = \frac{1}{(\mu - \lambda)}$$

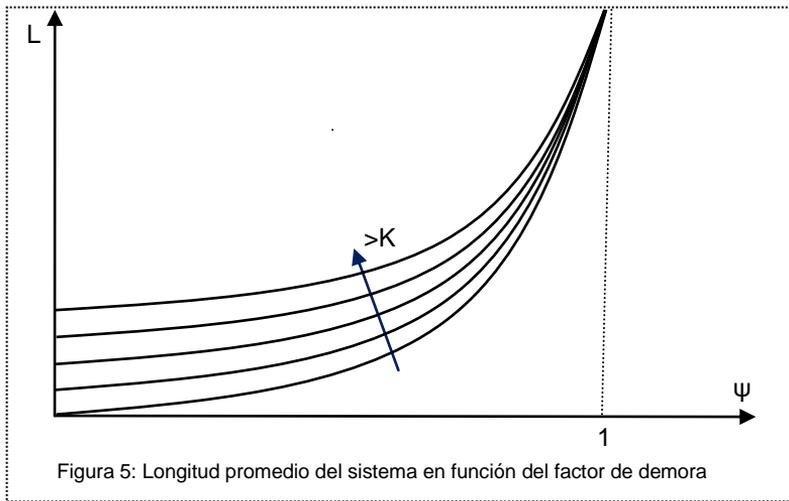
Si se reemplaza [25] en [29], operando se llega a:

$$W = \frac{1}{(1-\psi) \cdot \lambda} + \frac{K-3}{2 \cdot \lambda} \quad [30]$$

Aplicando Little una vez más, la expresión del número promedio de clientes en el sistema es:

$$L = \frac{1}{1-\psi} + \frac{K-3}{2} \quad [31]$$

La Figura 5 muestra esquemáticamente la variación de L en función del factor de demora para distintos tamaños de procesamiento.



Por otra parte, teniendo en cuenta [27], la longitud promedio de la línea de espera es:

$$L_C = \left( L - \frac{K-1}{2} \right) \cdot \frac{\rho}{K} \quad [32]$$

Además, el número promedio de clientes en la totalidad del sistema es igual a la suma de las cantidades promedio de los clientes que están esperando a que se complete el canal más los que están esperando en cola más los que están siendo atendidos en el canal:

$$L = B + L_C + H \quad [33]$$

Reemplazando [7], [31] y [32] en [33], y operando, se llega fácilmente a la siguiente expresión alternativa de B (más sencilla que [25]):

$$B = \frac{1-\rho_K}{1-\psi} + \frac{K-3}{2} + \rho_K - \rho \quad [34]$$

En definitiva, se han determinado expresiones para las variables de eficiencia más características de la instalación (fórmulas [30] a [34], [26] y [27]) en función de los parámetros del problema, sin necesidad de recurrir a las funciones generativas de probabilidades de estado de régimen permanente.

En el Anexo 1 se muestra el modelo P/P<sup>K</sup>/1 codificado en sistema LINGO<sup>5</sup> utilizando las fórmulas aquí determinadas. El usuario simplemente debe ingresar en la sección DATOS los valores de los parámetros K, LAMBDA (λ) y MU (μ).

### 2.5.5. Determinación del tamaño óptimo del grupo de procesamiento

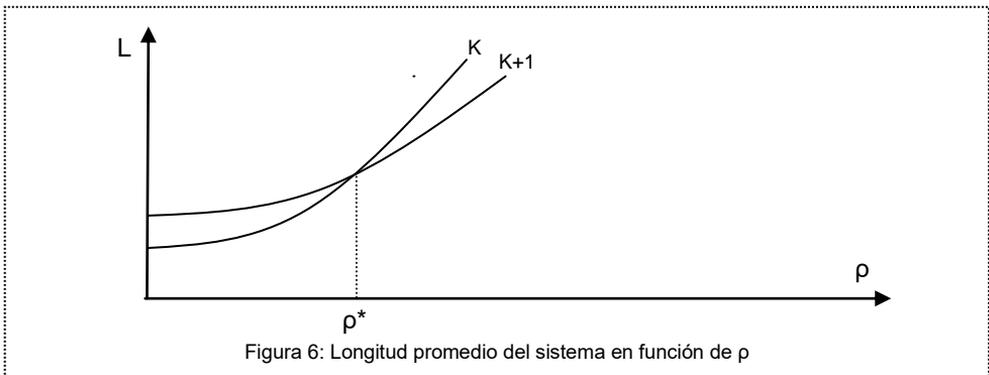
El dimensionamiento de la infraestructura que asegure un

<sup>5</sup> LINDO, marca registrada en USA de LINDO Systems, Inc.

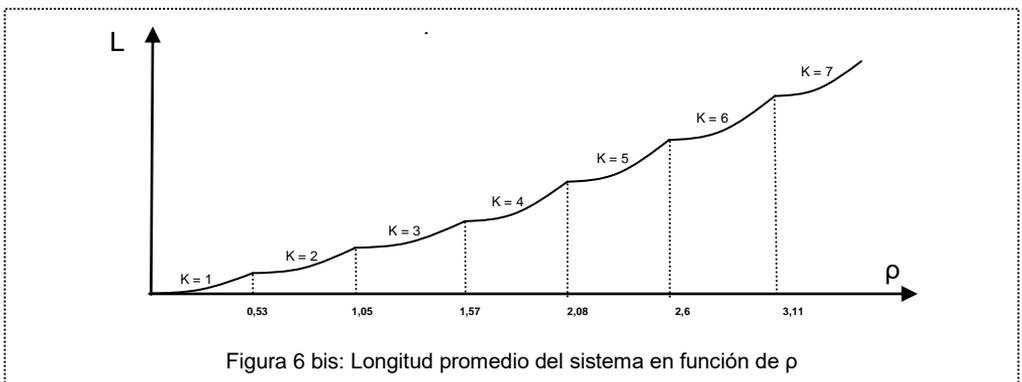
funcionamiento adecuado es una decisión importante que los administradores de los sistemas de colas deben tomar. En general, los usuarios perciben la eficiencia a través de las variables características, tales como  $L$ ,  $L_C$ ,  $W$  o  $W_C$ , mientras que los responsables del sistema lo hacen además mediante una función económica  $Z$ . La variable típica de decisión en los modelos  $P/P^k/1$ , es el tamaño  $K$  de los lotes a procesar.

Tal como se vio anteriormente en [8], la capacidad del canal para que el sistema se encuentre en equilibrio debe ser mayor al factor de tráfico, de manera que el tamaño mínimo del lote ( $K_{MIN}$ ), es igual al próximo entero superior a la relación  $\lambda/\mu$ . Por otro lado, a partir de las ecuaciones [17] y [31], la longitud promedio del sistema  $L$  es función de  $\rho$  y  $K$ .

Para un determinado valor de tráfico hay una capacidad que garantiza una longitud promedio mínima. Hasta un umbral  $\rho^*$  el mínimo valor de  $L$  se verifica en  $K_{MIN}$ , conforme muestra esquemáticamente la Figura 6, pero para un  $\rho$  mayor el mínimo se efectiviza para un tamaño  $K_{MIN}+1$  hasta un nuevo umbral.



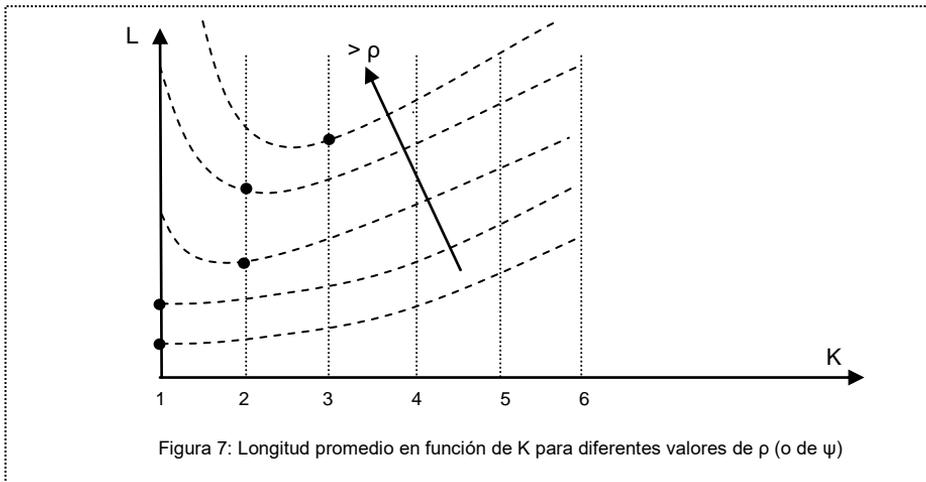
Habrán entonces rangos de valores del factor de tráfico dentro de los cuales hay una capacidad que minimiza la longitud de clientes en el sistema, tal como se indica en la figura 6bis.



En la Tabla 2 se muestran los tamaños óptimos para distintos umbrales

de  $\rho$  (como así también de  $\rho_K$  y  $\psi$ ) y las longitudes promedio correspondientes a ellos, para tamaños inferiores a diez unidades<sup>6</sup>. Por ejemplo, si el factor de tránsito de una instalación fuera igual a 2, el canal debería tener como mínimo una capacidad para procesar 3 clientes; sin embargo, la mínima longitud promedio de clientes se verifica para una capacidad de  $K = 4$ .

La Figura 7, en donde se representa esquemáticamente la variación de  $L$  en función de  $K$ , muestra adicionalmente el concepto del párrafo anterior. Para valores del factor de servicio menores a 0,527865, la función de  $L$  con respecto a  $K$  es siempre creciente. Para valores superiores, sin embargo, presenta una inflexión en el mínimo.



Se realizará a continuación un análisis económico a fin de determinar el tamaño óptimo de los lotes de procesamiento. Típicamente, los costos relacionados con los sistemas de servicio masivo son los siguientes:

- Costo de permanencia ( $c_L$ ), expresado en  $\$/(\text{cl}\cdot\text{t})$ : Es el costo en el cual se incurre por cada cliente que se encuentre dentro del sistema (por ejemplo, lucro cesante).
- Costo operativo ( $c_d$ ), expresado en  $\$/\text{lote}$ . Es el costo que se genera por procesar un lote, independiente de su tamaño (por ejemplo, costo de *set-up*).
- Costo de capacidad ( $c_K$ ), expresado en  $\$/(\text{cl}\cdot\text{t})$ . Es el costo de infraestructura del canal, es decir en el cual se incurre por cada cliente que el canal puede procesar simultáneamente (por ejemplo, alquiler, etc.).
- Costo directo ( $c_D$ ), expresado en  $\$/\text{cl}$ : Conjunto de erogaciones en las que se incurre con relación al servicio que se presta a cada cliente (por ejemplo, insumos).

<sup>6</sup> Umbrales mayores se pueden calcular igualando sucesivamente  $L(K)$  con  $L(K+1)$  de la expresión [31].

Luego, la expresión cuantitativa a minimizar, expresada en unidades monetarias por unidad de tiempo, es:

$$Z = c_L \cdot L + c_q \cdot \frac{\lambda}{K} + c_K \cdot K + c_D \cdot \lambda \rightarrow \text{Min}$$

Reemplazando el valor de L de [31]:

$$Z = c_L \cdot \left( \frac{1}{1-\psi} + \frac{K-3}{2} \right) + c_q \cdot \frac{\lambda}{K} + c_K \cdot K + c_D \cdot \lambda \rightarrow \text{Min} \quad [35]$$

En la Figura 8 se muestran esquemáticamente los costos intervinientes en función de K. Tal como se indicó anteriormente, la longitud promedio L con relación a K puede ser siempre creciente o tener un cambio de tendencia en un mínimo; y así será el primer sumando de Z. El segundo término es decreciente hiperbólicamente en función de la capacidad, el tercero es linealmente creciente y, finalmente, el último es una constante. En consecuencia, el costo total puede tener un mínimo único (como muestra la figura) o presentar un mínimo parcial y un mínimo global.

Para la determinación del K óptimo se recomienda utilizar un sistema de programación matemática con optimizadores discretos y no lineales. En el Anexo 2 se proporciona un programa expresado en LINGO para optimizar la función objetivo en base a los parámetros consignados en la sección DATOS.

Una aplicación muy interesante de este modelo de decisión es la determinación de la capacidad de transporte óptima para sistemas de inventarios aleatorios. En efecto, en depósitos de reaprovisionamiento constante (Miranda, 2006), en los cuales el suministro de un producto es permanente y la demanda es intermitente, ya sea instantánea o temporaria, siendo aleatorio por lo menos uno de los dos procesos, un objetivo habitual del problema es el dimensionamiento del medio de transporte que debe llevar el lote de la mercadería fabricada desde el depósito a otro destino (p.ej., al cliente), a fin de minimizar el costo total esperado. La analogía paramétrica entre el modelo de colas aquí descrito y el modelo de inventarios es la siguiente:

- L: Stock promedio (en depósito y en tránsito)
- H: Stock en tránsito
- $\lambda$ : Tasa promedio de fabricación del producto
- $T_S$ : Tiempo promedio que transcurre desde que el medio de transporte sale de planta hasta que regresa para una nueva carga
- K: Capacidad de carga del medio de transporte
- $c_L$ : Costo de mantenimiento en stock ( $c_1$ )
- $c_q$ : Costo de orden de transporte (k)
- $c_K$ : Costo por unidad transportada en el medio de transporte
- $c_D$ : Costo directo de fabricación (b)

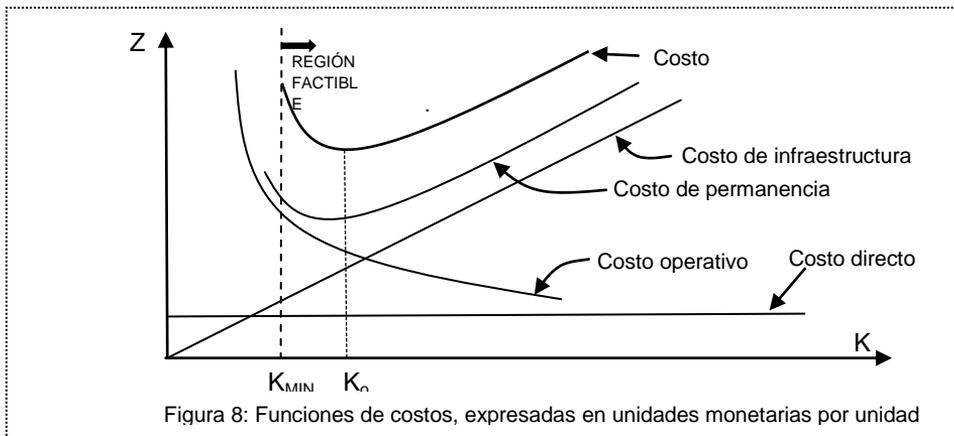


Figura 8: Funciones de costos, expresadas en unidades monetarias por unidad

### 3. CONCLUSIONES

En sistemas de procesamiento masivo de lotes fijos, la determinación de un “factor de demora” ( $\psi$ ) que contemple no sólo la relación del flujo ingresante con respecto al flujo saliente, sino también el tamaño del lote de procesamiento, permite calcular el número promedio de clientes esperando en el canal a que se complete un lote procesable (B). A partir de estas dos variables y de la consideración de conceptos substantivos de teoría de colas, en el punto 2.5.4 se propuso una metodología para calcular el número promedio de clientes en el sistema y en cola, como así también los tiempos promedios de permanencia y espera. Adicionalmente, el factor de demora permite determinar las probabilidades de estado en régimen permanente.

Se han definido rangos de los parámetros dentro de los cuales hay un tamaño de procesamiento que asegura el mínimo número promedio de usuarios en el sistema.

Asimismo, se ha determinado una expresión cuantitativa general para optimizar la capacidad de procesamiento del canal, y se han provisto códigos en lenguaje de programación matemática para los casos analizados. El modelo tiene múltiples aplicaciones prácticas en el ámbito de sistemas de producción de bienes y servicios, y en la administración de inventarios con demanda aleatoria.

Los sistemas de atención masiva están muy vinculados con los de arribos masivos. En efecto, en una red de colas, el lote de procesamiento de una estación puede constituir el lote que arriba a una estación subsiguiente. En consecuencia, el modelo es aplicable a redes complejas con estaciones de ingresos o egresos grupales, muy comunes en sistemas productivos. Si bien en este trabajo no se ha discutido la problemática de los sistemas masivos de lotes variables, la metodología aquí expuesta constituye un excelente punto de

partida para formular modelos analíticos que permitan determinar sus variables de eficiencia y su capacidad máxima de procesamiento.

### TABLAS Y ANEXOS

TABLA 1: Factores de demora en función del factor de servicio del canal y de la capacidad, en sistemas P/PK/1.

$\rho_k$	$\psi$											
	K = 1	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7	K = 8	K = 9	K = 10	K = 11	K = 12
0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.05	0.050000	0.091608	0.130688	0.166774	0.200051	0.230796	0.259274	0.285723	0.310351	0.333337	0.354841	0.375002
0.10	0.100000	0.170820	0.233038	0.287108	0.334263	0.375660	0.412256	0.444823	0.473985	0.500246	0.524014	0.545627
0.15	0.150000	0.241620	0.317339	0.379949	0.432301	0.476625	0.514594	0.547463	0.576184	0.601491	0.623956	0.644028
0.20	0.200000	0.306226	0.389365	0.455316	0.508660	0.552608	0.589402	0.620640	0.647482	0.670791	0.691217	0.709263
0.25	0.250000	0.366025	0.452536	0.518790	0.570952	0.613012	0.647615	0.676568	0.701142	0.722257	0.740593	0.756663
0.30	0.300000	0.421954	0.509017	0.573705	0.623501	0.662960	0.694974	0.721457	0.743723	0.762701	0.779068	0.793327
0.35	0.350000	0.474679	0.560256	0.622182	0.668949	0.705474	0.734771	0.758784	0.778821	0.795790	0.810346	0.822967
0.40	0.400000	0.524695	0.607272	0.665648	0.709011	0.742462	0.769038	0.790654	0.808578	0.823679	0.836575	0.847715
0.45	0.450000	0.572381	0.650805	0.705102	0.744854	0.775194	0.799101	0.818421	0.834356	0.847723	0.859096	0.868890
0.50	0.500000	0.618034	0.691414	0.741271	0.777307	0.804554	0.825871	0.843002	0.857067	0.868822	0.878791	0.887353
0.55	0.550000	0.661895	0.729528	0.774701	0.806977	0.831179	0.849996	0.865043	0.877348	0.887599	0.896269	0.903698
0.60	0.600000	0.704160	0.765485	0.805809	0.834323	0.855544	0.871952	0.885014	0.895660	0.904503	0.911964	0.918345
0.65	0.650000	0.744990	0.799557	0.834925	0.859697	0.878011	0.892098	0.903270	0.912347	0.919867	0.926199	0.931604
0.70	0.700000	0.784523	0.831967	0.862311	0.883380	0.898860	0.910714	0.920082	0.927671	0.933944	0.939216	0.943709
0.75	0.750000	0.822876	0.862898	0.888180	0.905593	0.918316	0.928018	0.935660	0.941836	0.946930	0.951203	0.954840
0.80	0.800000	0.860147	0.892503	0.912707	0.926519	0.936559	0.944185	0.950175	0.955004	0.958980	0.962310	0.965140
0.85	0.850000	0.896424	0.920913	0.936038	0.946308	0.953736	0.959359	0.963763	0.967307	0.970219	0.972654	0.974722
0.90	0.900000	0.931782	0.948237	0.958297	0.965083	0.969970	0.973657	0.976538	0.978850	0.980748	0.982333	0.983677
0.95	0.950000	0.966288	0.974572	0.979588	0.982951	0.985362	0.987176	0.988590	0.989723	0.990652	0.991426	0.992082
1.00	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

TABLA 2: Capacidad de procesamiento del canal que asegura la mínima longitud promedio de permanencia de clientes en el sistema en función de umbrales de parámetros. Longitud promedio correspondiente a dichos umbrales.

Valores a partir de los cuales K determina mínimo L			K	L <sub>i</sub>
$\rho$	$\rho_k$	$\psi$		
0,000000	0,000000	0,000000	1	0,000000
0,527865	0,263933	0,381967	2	1,118035
1,046994	0,348998	0,559274	3	2,268981
1,564852	0,391213	0,658324	4	3,426748
2,082314	0,416463	0,721223	5	4,587217
2,599546	0,433258	0,764647	6	5,748940
3,116704	0,445243	0,796395	7	6,911467
3,633750	0,454219	0,820608	8	8,074373
4,150797	0,461200	0,839684	9	9,237674
4,667774	0,466777	0,855096	10	10,40108

Programas del modelo P/P<sup>K</sup>/1 en lenguaje LINGO

ANEXO 1: Modelo descriptivo	ANEXO 2: Modelo optimizante
<pre> !DATOS; K =6; LAMBDA = 0.7; MU = 1.3 ; !MODELO; Ts = 1/MU; Ta = 1/LAMBDA; RO = LAMBDA/MU; RO = FI * (1-FI^K)/(1-FI); W = B*Ta + (L-(K-1)/2)*Ts/K + TS; L = 1/(1-FI) + (K-3)/2; W = L/LAMBDA; WC = (L-(K-1)/2)*Ts/K; L = W*LAMBDA; PO = (1-FI)/K; PA = RO/K; PI = 1 - PA; H = RO; LC = WC * LAMBDA; B = L - LC - H;                     </pre>	<pre> MIN = Z; !DATOS; LAMBDA = 23; MU = 10; CL = 10; CQ = 2; CK = 30; CD = 0.5; !MODELO; RO = FI * (1-FI^K)/(1-FI); L = 1/(1-FI) + (K-3)/2; H = RO; RO = LAMBDA/MU; @GIN(K); Z = CL * (1/(1-FI)+(K-3)/2) + CQ*     LAMBDA/K + CK*K + CD* LAMBDA;                     </pre>

**REFERENCIAS BIBLIOGRÁFICAS**

BOLCH, G.; GREINER, S.; DE MEER, H.; TRIVEDI, K. S. (2006): "QUEUEING NETWORKS AND MARKOV CHAINS". Editorial: John Wiley & Sons, Inc., USA.

GROSS, DONALD; HARRIS, CARL M. (1988): "FUNDAMENTALS OF QUEUEING THEORY". Editorial: John Wiley & Sons, Inc, USA.

LITTLE, J. D. C. (1961), "A PROOF FOR THE QUEUEING FORMULA:  $L = \lambda W$ ", Operations Research, Vol. 9, No.3, pgs.383,387

MIRANDA MIGUEL (2013), "TEORÍA DE COLAS", Editorial Educa. Buenos Aires.

MIRANDA MIGUEL (2006), "SISTEMAS DE OPTIMIZACIÓN DE STOCKS", Editorial Educa. Buenos Aires.

WHIT, W., MELAMED B. (1990), "ON ARRIVALS THAT SEE TIME AVERAGES". Operations Research, vol. 38, No. 1, 1990, pgs. 157, 172.